NAVAL RESEARCH LOGISTICS QUARTERLY

OFFICE OF NAVAL RESEARCH

CONTENTS

ARTICLES	Page
Some General Observations and Experiences in Logistics *Admiral Robt. B. Carney (Retired)*	1
System Analysis and/or Common Sense *Commander Yngve Rollof (Royal Swedish Navy)*	11
The Dynamics of Overhaul and Replenishment Systems for Large Equipments *Joseph Fennell and Seiki Oshiro*	19
The Concept of A "Mission" *Michel Verhulst*	45
Various Optimizers for Single-Stage Production *Wayne E. Smith*	59
Notes on the Theory of Dynamic Programming. IV – Maximization over Discrete Sets *Richard Bellman*	67
Attrition Games *J. R. Isbell and W. H. Marlow*	71
An Algorithm for Quadratic Programming *Marguerite Frank and Philip Wolfe*	95
The Optimization of A Quadratic Function Subject to Linear Constraints *Harry Markowitz*	111
NEWS AND MEMORANDA	135
RECENT PUBLICATIONS	137

Bind in

1956-1957

Soc R v.3-4

AUG 20 1956

NOV 18 1960

VOL. 3, NOS. 1 and 2

MARCH and JUNE 1956

NAVEXOS P-1278

NAVAL RESEARCH LOGISTICS QUARTERLY

EDITORS

Rear Admiral H. E. Eccles, USN (Retired)
The George Washington University

Commander G. S. Essenwine, USN
Office of Naval Research

O. Morgenstern
Princeton University

F. D. Rigby
Office of Naval Research

M. E. Rose, *Managing Editor*
Office of Naval Research
Washington 25, D. C.

ASSOCIATE EDITORS

E. W. Cannon, *National Bureau of Standards*
T. B. Evans, *Colonel, USA*
P. L. Folsom, *Captain, SC, USN*
M. A. Geisler, *The Rand Corporation*
J. D. Hayes, *Rear Admiral, USN (Retired)*
E. B. Henry, Jr., *Commander, USN*
A. J. Hoffman, *National Bureau of Standards*
S. Karlin, *California Institute of Technology*
J. Laderman, *Office of Naval Research Br. Office*
A. T. Magnell, *Captain, SC, USN*
W. H. Marlow, *The George Washington University*
J. Marschak, *Cowles Foundation*
T. C. Roberts, *Commander, SC, USN*
H. P. Sachaklian, *Colonel, USAF*
E. K. Scofield, *Commander, SC, USN*
J. S. Skoczylas, *Lt. Colonel, USMC*
E. D. Stanley, *Captain, SC, USN*
R. M. Thrall, *University of Michigan*
C. B. Tompkins, *University of California*
A. W. Tucker, *Princeton University*
R. E. Williams, *Commander, SC, USN*
M. Wood, *Office of the Secretary of Defense*
M. A. Woodbury, *The George Washington University*

The Naval Research Logistics Quarterly is devoted to the dissemination of scientific information in logistics and will publish research and expository papers, including those in certain areas of mathematics, statistics, and economics, relevant to the over-all effort to improve the efficiency and effectiveness of logistics operations.

Manuscripts submitted for publication should be typewritten, double-spaced, and the author should retain a copy. Manuscripts and other items for publication should be sent to the Managing Editor at the above address. Authors will receive 50 free reprints of their published papers.

The Naval Research Logistics Quarterly is published by the Office of Naval Research in the months of March, June, September, and December and can be purchased from the Superintendent of Documents, U. S. Government Printing Office, Washington 25, D. C. Subscription Price: $1.50 a year in the U. S. and Canada, $2.00 elsewhere; 50¢ for a single copy. Reprints of an individual article can be purchased in quantities of 100 or more. Requests for the purchase price of reprints of a particular article should be sent to the Superintendent of Documents.

The views and opinions expressed in the articles in this Quarterly are those of the authors and not necessarily those of the Office of Naval Research.

The printing of this publication was approved by the Director of the Bureau of the Budget, July 26, 1955

SOME GENERAL OBSERVATIONS AND EXPERIENCES IN LOGISTICS

Robt. B. Carney, Admiral USN (Ret.)
The George Washington University Logistics Research Project

The author, recently retired after a distinguished Naval career, discusses the impact of logistics on high-level planning.

This dissertation on logistics represents my personal views and reflects experiences and conclusions reached over a period of years; from time to time, my own conclusions underwent alterations as added experience broadened my views and understanding. First, a few observations concerning my experiences in the logistical field in World War II might be of interest.

In February 1941 I was ordered to report to the Navy Department where I found an atmosphere of secrecy and mystery concerning my own assignment. Eventually I reported to Rear Admiral Bristol who had been designated the "Commander Support Force." That this name was deliberately misleading will be understood when I say that the mission of the Support Force was to organize, equip, and train a force for early participation in the protection of shipping. Mind you, this was in February 1941—ten months before Pearl Harbor. Briefly, the concept of operation was the establishment of a force which would operate from bases in the North Atlantic for the protection of shipping by the escort of convoy, by certain air operations, and by port protection. The Navy was in no respect ready, material-wise, to put such a concept in operation, and even a cursory study of the problem pointed out clearly that we would need such things as tank farms, hospitals, water-front facilities, ship repair units, air repair units, and a host of other things involving material not on the shelf and personnel as yet untrained.

The problem itself at least indicated an organizational solution, which eventually resulted in my being designated as Officer in Charge of Plans and Operations—perhaps the first time this combination had ever been employed in the Navy.

For many years the Navy had been so obsessed with the tactical aspects of the Battle of Jutland that little or no attention had been paid to the thinking and the planning that was involved in a sustained campaign wherein losses were affected by incremental attrition, rather than by the holocaust of the major fleet engagement. Consequently, the Support Force staff started from scratch in its planning and its accumulation of equipment.

It is of interest to note that when the Japs struck at Pearl Harbor on the 7th of December 1941 there was no accumulation of equipment for the establishment of an advanced base in the Western Pacific and, in order to fit out the first base at Bora Bora (known as "Bob Cat" in the code parlance), it was necessary to commandeer the base-development equipment which had been accumulated by the Support Force for the building of its North Atlantic bases. It also is of interest to note that the Support Force representative in Washington, Captain W. A. Corn, had worked out a sort of catalog of components for advanced bases — the first document of its kind in the Navy.

For a time I had a respite from these matters while commanding the USS DENVER. Then, in July of 1943, I was ordered to report to Admiral Halsey as his Chief of Staff. At that time the most significant deficiency in the entire organization of the South Pacific was lack of

adequate coordination of logistical planning and logistical control to conform to the requirements of the approved plans, and to the support and maintenance of the operating forces which frequently were engaged in surprise operations for which no plans had been made nor could have been made. It was here that the full impact of the importance of logistical planning from top to bottom became apparent to me. We set about developing a carefully integrated system of headquarters planning and control which would marry strategic and operational requirements with the realities of logistics; to a reasonable extent we accomplished this marriage within the South Pacific area, but it became apparent to us that what happened in the Joint Chiefs of Staff, and what happened in the various civilian agencies for control of our industry and resources, all had a dynamic effect—or perhaps a negative effect at times—on the affairs of the South Pacific.

Later, when Admiral Halsey went to command the operations of the Third Fleet in the Central Pacific, our attention was turned to mobile logistical support, and I received a good education in the floating advanced base concept and the mobile logistics support attending the forces at sea.

Although I spent from June 1940 to December 1945 with the operating forces, much of it with the combat forces, I returned to the Navy Department after the war with a profound conviction that line officers had to get into the logistical business, get into it quickly, and get into it in a big way. I sought appointment as Deputy CNO for Logistics and welcomed the opportunity to tackle that job even though I thoroughly recognized my own inadequacy.

One of the first things which came to my attention was the almost complete lack of senior officers who were broadly qualified in the philosophy, concept, and operations of the logistics business. They could virtually be named on the fingers of one hand; there was, of course, Admiral Horne, who had evolved as the virtual coordinator of the entire logistical operation of the Navy Department in support of the operating forces; there was Admiral Badger, who had been associated with the beginnings of a systematic organization for the performance of the logistical function in the form of what was then known as OP-12; there was Admiral (then Captain) McCormick, who had also served in OP-12. There were certain officers who had specialized experience in personnel, in the specialized aspects of aviation, and in other restricted fields; there also were officers who had served with the Service Forces, particularly in the Pacific, who had a keen understanding of the requirements in end-products but had not rubbed shoulders with the shore establishment aspects nor the problems of procurement. Considering that the best operational plan in the world is no better than the logistical support that it receives, and considering the vast material implications of the Navy's expansion and the campaigns that it had waged, it is extraordinary that there had been so little appreciation among the line personnel of the vital importance of the great science of logistics. Nevertheless, the dearth of qualified personnel was a real one.

Early in 1946 I set about organizing in OPNAV for the performance of the various logistical functions for which the operating people were responsible, and I placed particular emphasis on obtaining and training senior Captains and Flag Officers in these responsibilities. At the end of four years, by my own reckoning I had put about 18 Flag Officers through the mill so that the Navy was in far better position in that respect than it had been at the end of the war. With a reservoir of 18 Flag Officers and a great many highly competent Captains who had served in OP-04, the operational future, logistics-wise, was assured.

It was also necessary to emphasize the fact that the selection for promotion process was not ignoring those who had served well in the logistical fields; a consistent effort was made

to impress the importance of logistics on all hands to the end that the Selection Board members would regard logistics as an essential function of operations and would give careful consideration to those who had performed well in the logistical field. I consider that a "working competence" in logistics is a requisite part of the professional competence of all unrestricted line officers. I consider it a requirement for high command, and I also consider it a requirement, in appropriate measure, in the lower ranks. The results of this effort were reflected in an increasing number of "logisticians" being selected to Flag rank. I felt that my own selection as CINCNELM and subsequent appointment as CINCSOUTH and Chief of Naval Operations should have served as an encouragement to those who performed in the logistics bailwick.

Next, let me say that the word "logistics" seem to have sprung into being, full-grown, from the dragon's teeth of the urgent needs of war. Curiously enough — right up until now — there is nothing which even approaches a universal understanding of the full meaning of "logistics;" this is true in the Navy and throughout the entire governmental structure. There is a formidable body of opinion which regards "logistics" and "supply" as interchangeable, and many otherwise well-informed military men still regard logistics as something which should be sloughed off onto the technical people and the commercial experts. Among the operating personnel there is still a lingering suspicion of logistics and some feeling that to be identified with logistics is to receive the kiss of death. Actually, logistics is an all-hands responsibility, and no officer who aspires to high naval operating command can be considered properly qualified for such responsibility if he does not have a firm basic understanding of the importance of logistics and the organization by which logistics is administered.

Logistics is a subject on which volumes could be written, the more so because there is a persistent unwillingness, or inability, to place "logistics" and "supply" in proper mutual relation and proper perspective. The error has been compounded and perpetuated in the designation of the Office of the Assistant Secretary of Defense for Supply and Logistics; that title is a misnomer and has the tail wagging the dog.

The distinction between logistics and supply is implicit in a definition evolved by the George Washington University Logistics Research Project: "the process of planning for and providing goods and services." Logistics is the entire system of planning for and providing goods and services; it is a campaign which begins with the statement of requirements, proceeds through the business of procuring and only ends with final distribution. Perhaps one might say that it does not even end there, because there might be storage and preservation problems at the very end of the line rather than utilization in the sense of being consumed or stricken from the inventory. Aspects of logistics include personnel, transportation, supply, medical, maintenance and repair, construction, storage and other specialties, but they all are part of the big logistics picture. In addition, the solution of a logistic problem has the management aspects of organization, planning, execution and supervision. The whole process calls for the efforts of the operational forces, the military men of all groups and corps, the shore establishments, participation by civilian executives and technical people in the departments, and the best brains of industrial management, design, and production. Regarded in this light, it is perfectly evident that "supply" is but one facet of the greater function known as "logistics."

An inspection of the sequence of events involved in preparing for any military operation of whatever size clearly points up the fact that it is an all-hands problem. Be it a small patrol operation or a major campaign, two things must be placed in balance before the operation can be undertaken with any hope of success: objectives and feasibility. The military planner outlines those things which he would like to accomplish—his objectives; these objectives are then

tested against the capability of our side to produce the men and material initially needed, and to produce the replacement men and material required for the prosecution of the operation to its ultimate success. The establishment of this balance is imperative, and procedures must be established at every level for its accomplishment.

I have always visualized naval logistics as a dynamic flow system, not only of men, goods, services, etc., but as a planning and executive effort. This thought did not originate with me, I am sure, because I found it, piecemeal, in the minds of many people as I studied the problem immediately after the war. But I may have rendered some service to the cause of logistics by pulling the variously related ideas together and preaching them as one system, or flow.

As a distinction between the purely military phases of logistical planning and those steps which must be performed by technical personnel in the military establishment and by the industrial suppliers, the expressions "producer logistics" and "consumer logistics" have come into common use. On close examination, however, the process does not lend itself too well to division into two clearly defined and separate processes. Rather, the flow of events resolves itself into three phases: (1) the statement of requirements, (2) procurement, and (3) distribution. There is some measure of fuzziness even in this division, but it appears to better point up the actual steps.

The statement of requirements is primarily a military function—tabulating those things which are needed to successfully carry out the strategical and technical plans. Procurement is the process which involves a translating of military requirements into technical specifications and the actual production of those things which are needed, either by industry or by the facilities of the Department of Defense. Distribution is a process which begins when items are completed and ready for use; it may be considered to begin at the end of the procurement phase but ultimately laps over into the military operation. In the case of fuel, the final disposition is up the stack; in the case of bullets, the final stage of distribution is through the muzzle. From this it can be seen that the operating personnel are vitally concerned with the first and third phases which therefore automatically become essentials in the operating business.

As stated before, logistics is a subject of direct concern from the smallest field unit right up to the national government level. Each level has its own problems of determining objectives and feasibility; although differing in degree, each level must concern itself with a statement of requirements, procuring the things that are needed, and distributing them in conformance with the operating plans concerned. This had always been recognized to some extent by the operating forces and in the individual departments, but in the security law that was passed in 1947 the drafters took cognizance of the fact that there was a national level at which strategic and logistical planning must be carried on to the end that national policies and national objectives are feasible in terms of the national resources available. This was clearly indicated by the creation of the National Security Council and the National Security Resources Board.

Had the National Security Resources Board developed as envisaged by the drafters of the 1947 law, the government would have had, at the top level, an agency for handling the national logistical problems in terms of its major human and material resources. Actually, extraneous political factors prevented fruition of the plans for the National Security Resources Board, and it atrophied to the point of being disestablished; instead, the Munitions Board of the Department of Defense initially performed many of the functions which should have been undertaken at the national level rather than the level of the Department of Defense. Later, the national logistical functions was placed in the hands of the Office of Defense Mobilization.

There was still one more level which had to be considered; I pointed this out in a speech in May of 1950. Ultimately it would be necessary for the freedom-loving nations to band together in their own mutual defense and, at that time, in addition to the problems of grand strategy, it would be necessary to enter into arrangements for considering the "grand logistics" of alliances. In other words, there was in the offing an international logistical problem. This all came to pass in the creation of NATO.

In summary, it is essential that officers assigned to the high planning, administrative, and command positions have a working understanding of the economic-logistic relationship at the top military levels and the lateral and higher levels in other agencies of the government concerned. Otherwise, the Navy will lose its shirt, and perhaps its pants as well.

Before proceeding to the planning system, it would be well to mention briefly the impact of "unification" on the business of logistics. Late in the war, great momentum was generated for a philosophy that the Services should be "unified." No one knew just exactly what "unification" meant, but certain of the claims of its proponents had powerful appeal and secured a tremendous weight of public opinion support. Because the proposal was not without great danger to certain fundamental principles to which the Navy adhered, the Navy in general strongly opposed the general philosophy of unification; however, even at that time, I considered that the Navy's views were as negatively extreme as were the views of the Army and the Air Force and certain civilian elements extreme in the direction of merger. That battle is history and need not be reviewed here except to say that the profligate expenditure of the nation's human and material resources during the war was such that sober thought should certainly have dictated the adoption of whatever measure of consolidation could be achieved without detriment to the efficiency and flexibility of the operations of the fighting forces. Perhaps the pendulum has swung too far in the direction of "merger" with respect to logistics, but on the other hand much has been accomplished under the unified direction of the Department of Defense which could never have been achieved solely on the basis of obtaining voluntary agreements and voluntary submerging of service autonomies. In any event, a pattern of top level control at the level of the Department of Defense has been established; the Services must accommodate themselves to that pattern even though they should, I believe, strongly resist the current tendency to reach down into segments of the several departments without reference to the men responsible in those departments, and thus by-pass the proper authoritative chain of command; this tendency is real, it is extensive, and it could be fatal in time of crisis.

Much has been said in the past 15 years concerning joint service procurement and its ramifications. First, I would like to emphatically veto any proposal for a fourth Service of Procurement. This idea has reared its head in many different forms over the past decade, and I am afraid that in certain categories of procurement it has already come to pass in the vesting of over-riding authority in certain divisions of the Department of Defense. Nevertheless, it is fundamentally unsound to place control of the quantitative, qualitative, and priority aspects of procurement in the hands of a Service not responsible for the outcome of military campaigns. For a long time important elements of the Army and Air Force supported the philosophy of a fourth Service of Supply, but for the most part they seem to have come to see the picture in the same light as viewed by the Navy.

There may be some merit in single service procurement in which one Service is designated the responsibility for procurement for all three other Services. However, it is expecting too much to believe that one Service could altruistically accord priority to a sister Service o. an item which enjoyed high priority in its own list of needs; to some extent this factor of human

weakness can be overcome through the use of a working fund to which contributions are made by each Service and against which withdrawals can be made while credit exists. This device is employed with reasonable success in the operation of MSTS, each Service being able to order services as long as its credit is good in the working fund. It is my own belief that certain items can best be procured for the Department of Defense by a single Service provided that (a) design and specifications are satisfactory to all users and (b) that each Service budgets for its own needs, placing appropriate funds at the disposal of the Service designated for the single procurement job.

I would like now to speak concerning the planning system. I believe that the planning system which is in effect in the Navy Department is essentially sound and sufficiently flexible to accommodate itself to changing conditions. Detailed information on this planning system is readily available and need not be covered here except to mention two very important aspects of the philosophy of the system.

Before the present system was first put into effect, there was one defect which frequently manifested itself: there was a tendency on the part of each planner, at each stage, to insist on the completion of his own particular step in the planning sequence before releasing his efforts to those in charge of the next step. This entailed unacceptable delays—delays so great that by the time the last step was completed those in charge of the first step were busily engaged in revision. In other words, the production of a completed plan was virtually impossible. This led to the concept of "concurrent planning" which I conceived and put into operation in the Navy Department planning system. The principal of concurrent planning is simply one of keeping the planners of the preceding and successive phases informed as to the development of the plans to the end that, by continuing consultation, the best thinking on cause and effect is continually available in the development of each step in the planning sequence.

The other concept which is worthy of note is that of the General Planning Group. The General Planning Group, functioning directly under the Vice Chief of Naval Operations, having representation from every activity in the department, provides a clearing house, at the working level, for planning. It not only is of inestimable help in the preparation of strategic and logistical plans, but it also serves as the focal point for the developing of the budgetary plans by which we tailor an optimum naval establishment within the means at our disposal.

Based on my experiences with logistical planning in the field during the war, and on the reviews that I made when becoming Deputy Chief of Naval Operations for Logistics after the war, the inevitable conclusion was that there had been considerable excessive procurement; this, taken into consideration with the realization that the United States could not afford to repeat such profligacy, pointed to the urgent need for a system of logistical planning which, while providing adequately, would obviate waste of our precious resources. The programming for a campaign when conducted manually involved unacceptable delays; the introduction of business machines into the process was an improvement but still did not meet the demonstrated requirements for speed, nor could they adequately integrate all of the numerous and complex factors which had a bearing on the calculations and the final figures. Obviously, the problem was now too complex for the human mind to comprehend and handle within a reasonable time; therefore, some entirely different system was required to meet the demands of speed and conservation. These inexorable facts and conclusions led me to believe that recourse must be had to the capabilities of electronic computation which were even then under study at the Special Devices Center and at various laboratories.

Having come to that conclusion, I proceeded to mobilize thinking on the subject; the response was spotty because there was not universal grasp of the magnitude and gravity of the problem, nor did any of us see how the factors of a problem could be fed into a machine nor how the reaction to the machine could be interpreted in concrete terms. Nevertheless, largely as a result of interest by the Office of Naval Research, a general conference was set up for the purpose of discussing the problem and the possibilities for its solution. This conference was attended by operational and technical people interested in logistics, electronic experts, mathematicians, representatives from the Bureau of Labor Statistics, representatives from the Department of Commerce, and representation from every agency having a natural interest in any of the three phases of logistics, namely, military requirements, procurement (including national resources), and distribution. These discussions produced unanimous agreement among the scientific personnel that the construction of an appropriate electronic computer was feasible.

Such was the genesis of the Logistics Research Project set up at The George Washington University, and it was from these beginnings that the ONR Logistics Computer evolved. Mind you the above conference took place in the Spring of 1948, and since then many types of general and special-purpose, high-speed, electronic computers have been developed and have become commercially available.

High-speed electronic computers are relatively a new tool of wide applicability. Their utility is, of course, important to the military for at least two reasons:

(a) Many service problems, both operational and logistical, involve the handling of large masses of data routinely.

(b) Many unsolved military problems which can be subjected with profit to scientific research involve in their solution the handling of complex mathematics which can be most readily manipulated by modern computers.

Computers are by no means confined in their application to service matters. It is true that initially their use seems to have been generated by emergency needs, but as their utility developed their application to other fields also developed rapidly. The specific Logistics Computer was developed to serve the particular need for a planning tool to handle voluminous data and for time-phasing logistical requirements. It is a good example of "specialization" in the computer field.

The present have-not position of our country with reference to critical materials and qualified manpower is serious. These shortages are being continually accentuated in ever greater degree by reason of the complexity of the weapons and equipments of our era. Profligacy in the utilization of scarce material or the employment of highly skilled manpower required to fabricate the modern sinews of war is so serious as to have connotations of early defeat — it is as serious as that. This fact, coupled with the demonstrable inability of manual and mechanical machine methods to handle modern complex programming, points inevitably to the need of something better, and the only obtainable improvement appears to be through the medium of electronic computation.

With the introduction of atomic weapons, thinking went hay-wire for a time. As the possibility of the tactical use of atomic weapons dawned on military and other thinkers there was a further flurry of wishful and incorrect thinking. It has been very difficult to dispel the idea that the push button will decrease the logistical requirements of war; and yet early studies which I conducted while commanding the Southern Forces of NATO indicated to me that the introduction of tactical atomic weapons would increase the logistical demands on ground forces

by reason of the vastly increased requirements for dispersal and mobility. In other words, overhead would be increased by dispersal; transportation requirements would be vastly increased; and the maintenance required for increases of transportation equipment, increases in the number of depots and dumps, and the vastly increased number of highly trained technical personnel would all throw an even greater logistical burden on the ground forces. What was true of the ground forces also was true of the air forces. And now the naval forces are beginning to feel the impact of that problem.

Atomic weapons, atomic powered planes, and atomic powered ships all radically and seriously boost the cost of those weapons, planes, and ships. Perhaps the new problems introduced by the advent of utilizing the atom may not be new in the mathematical sense nor in the required functioning of the machine, but it is certain that many new factors must be considered in preparing military forces for combat operations, for sustaining them in combat operations, and for the prosecution of major campaigns. These complexities only emphasize the need for the development of essential data and for the feeding of that data into high-speed computers.

Here again I consider it imperative that the attention of top-level leaders be focused on the potential value of high-speed computers to the end that they can exert the necessary pressure and guidance to insure proper interest and support at the working levels.

I now come to my last point: the logistical aspect of NATO and ANZUS. I think that it would be safe to say that we could disregard the ANZUS for the time being; the nature of that treaty is such that there is a clear responsibility for the forces to be logistically supported by their own national organizations. Whatever might be gained by mutual arrangement would, of course, be so much velvet. For example, port facilities, provisions, etc., might be furnished by our Pacific neighbors, while we in turn furnished oil for the common use and performed such other services as were within the capabilities of our own existing forces and logistical support. In other words, the ANZUS organization is a loose one, and although there has been a certain amount of continuing planning, the nature of the geography and the composition of the forces that would be involved are such as to minimize the current importance of the logistical problems from an alliance standpoint.

In NATO the situation is quite different. There are several factors which enter into the problem, and these factors can at least be named, although it is not practicable to elaborate here. Although there is a tacit understanding that each nation will furnish that which it is best able to furnish, there is also a conflicting conviction on the part of the partners that individually they cannot afford to completely neglect any component, for to do so would be to destroy the national balance which each would like to maintain. The maintaining of balanced forces within each national structure is obviously impracticable because of the costs involved, and it is equally obvious that special contributions should be made by certain countries that are capable of making them. For example, the small nations must rely on the United States primarily for strategic air support. The same is probably true for tactical air support although some of the larger partners can make their contributions. On the other hand, we expect our European partners to supply the bulk of ground forces numerically.

Next comes the problem of standardization. Efforts have been made in this direction, and are continuing; but as long as there is a heterogenous collection of weapons in the NATO arsenals, the problems of resupply are in some cases virtually impossible of solution.

Then there is the question of stockpiling. Frankly, there is not sufficient ammunition and equipment stockpiled to permit continuation of combat beyond a certain point — and that

point comes uncomfortably soon after D-Day. What then is the answer? Increasing stockpiles overseas is an easy and obvious answer, at least statistically; however, there are military disadvantages as well as budgetary hurdles which make such a program virtually impossible of accomplishment.

Next comes the ever-present specter of budget. Increases of military expenditures have been dynamite in the European countries since before the meeting at Lisbon in early 1952; the politicians are unwilling to raise the ante in their own countries. On the other hand, people of the United States have become increasingly reluctant to provide military aid on items possibly within the capabilities of our neighbors and allies to supply. The result has been an impasse which affects NATO military readiness adversely and seriously.

The net result of all of this, in NATO, is that the factors of the problems are well known, the problems can be accurately stated, but the solutions to many major problems and their components remain unsolved and very probably many will continue unsolved.

The determination of requirements has been well handled, and I believe that the requirements are accurately known and have been accurately compiled and promulgated. The procurement phase of NATO logistics, if not bogged down, is at least very materially slowed down. Procurement is a complicated process dependent upon negotiations which invariably have a political flavor and are profoundly affected by international arrangements of an economic nature; strategic distribution presents few problems at this time; distribution in the communications zone is again affected by many political considerations; distribution in the combat zone must inevitably come under the combat commanders although at present there are political restrictions which are only waived reluctantly during actual maneuvers.

In conclusion, I desire to emphasize that logistics is the entire system of planning for and providing goods and services for the support of the Military forces. It is an all-hands maneuver. The whole process calls for the efforts of the operational forces, the military men of all groups and corps, and the shore establishment, participation by civilian executives and technical people in the departments, and the best brains of industrial management, design, and production.

Supply is only one phase of logistics. By placing "logistics" and "supply" in proper mutual relation and proper prospective, a more universal and firm basic understanding of the importance and scope of logistics will ensue.

Logistics is an essential component in the qualification of officers administering top command and the important segments of the naval organization.

The present Navy planning system is essentially sound and sufficiently flexible to accommodate itself to changing conditions. It is an integrated system of planning and control, which marries strategic requirements with the realities of logistics.

I would like to emphatically veto any proposal for a fourth Service. It is fundamentally unsound to place control of the quantitative, qualitative, and priority aspects of procurement in the hands of a Service not responsible for the outcome of military campaigns.

I consider it imperative that the attention of top-level leaders be focused on the potential value of logistics research, and particularly to high-speed computers, to the end that they can exert the necessary pressures and guidance to insure proper interest and support at the working levels. This is the only answer to the present have-not position of our country in respect to shortages in critical materials and qualified manpower.

* * *

SYSTEM ANALYSIS AND/OR COMMON SENSE

Commander Yngve Rollof
Royal Swedish Navy

The author discusses the rôle of scientific methods in the design and evaluation of military systems.

Nowadays there are often many different possible solutions to a technical military problem. Therefore, systematical methods have been developed which can be employed to solve such problems analytically and to establish technical data for weapons which will be optimal should the problem situations arise in real life. In order to be able to group together the requirements of a weapon, it is necessary that the study be based on a correct evaluation in which exact technical analysis is linked together with sound judgment and common sense. The essence of the procedure is to reduce the uncertain factors as far as possible by means of scientific systematics and due respect for facts. In the long run, it is profitable to expend considerable competent effort on an unbiased analysis instead of possibly making ill advised use of existing limited construction capacity for weapons that subsequently may be of limited value. The time is past when improvisation can save what is lacking in foresight!

The concept, system analysis, implies that a weapon system should be seen as a whole and not as a collection of different units. For instance, the tank, the destroyer, or the fighter is a weapon carrier, and its separate fighting equipment must always be considered in relationship to all the other fighting units. In short, the analysis is raised to a higher level than that applicable to an isolated weapon or vehicle. System analysis, like other forms of operations research, gives information about the measures leading to the best result under specified circumstances. An analysis of causes, reasons, and connections serves to replace subjective opinions with accurate facts. A scientific operations analysis is made by means of different mathematical methods. Although advanced mathematical methods may be required, many problems can be solved by very elementary mathematics.

The procedure for system analysis cannot be generalized. Every problem requires its special methods of solution. However, an attempt is made below to describe a principal procedure for system analysis.

PRIMARY VALUES

First, all available information about the problem must be collected. However clever is the mathematics used, the result of a mathematical analysis is never more accurate than its basic figures. One never gets a better result by using more decimals than are justified by the underlying data. As the German expression goes: "Die Genauigkeit kommt ja erst mit dem Logarithmieren." One must try to judge the reliability of the basic data and also to estimate the limits which it imposes on the accuracy of the mathematical values derived from them. The quality of the information determines whether or not the analysis will be correct. Information is the foundation of the entire process, and if the foundation is insecure the whole building will be insecure. It is therefore necessary to use an "information filter," deciding what is

incomplete, unreliable, subjective, or not applicable. It is a matter of producing actual facts, verified and complete facts. Especially, it is necessary to verify that information is up-to-date and not too old. In case he does not take extraordinary pains to evaluate the basic figures, even the most clever person will work with poor efficiency. This cannot be insisted upon too strongly. Many persons are of the wrong belief that extensive statistical techniques can compensate for low quality of data. It is certainly true that perfect techniques may squeeze the last drop of information out of given facts; but they can not compensate for lack of information nor eliminate faulty information. Instead, a far-reaching analysis requires an exceedingly high quality of primary values.

The collecting of exact information requires at least as much intelligence, far-sightedness, and imagination as it does calculations. Especially, military judgments have a tendency to be categorical and tendentious. Partly, this may depend on the military education, which requires a "resolute" behaviour and short, clear answers to correspondingly worded questions. Facts behind figures are, ordinarily, full of relative proportions, means, compromises, advantages versus disadvantages, etc. However, in order to emphasize the words, military facts are generally presented in a categorical manner. One favours a cause whole-heartedly or he is just as whole-heartedly against it. To the scientist seeking after truth this is one of the greatest difficulties when working with military matters. The following example shows how important it may be to examine available data: In the beginning of the Second World War the British Admiralty asked that an anti-aircraft gun should be designed that admitted a vertical fire for fighting. Crews on ships attacked by German Stuka dive bombers had reported that these planes did attack vertically. Investigations showed that "vertically" was, as a matter of fact, 45° - 75° from the horizontal plane, which makes a great difference to the designer of weapons. Thanks to these investigations, far-reaching constructional work and manufacture of new complicated material was avoided. If the information base is increased enough, the decision very often comes automatically.

DEFINITION OF THE PROBLEM

It is necessary to have a "Frenchman's attitude" to clearness and exactness. It is not sufficient just "to feel what you mean." The problem has to be defined and worded clearly, which sounds very easy but is in reality very difficult. Sometimes, the problem is solved in the same moment as it gets its exact wording. One must ask himself: "To what do I really want an answer?" A classic example of such a question concerned the Atlantic convoys during the Second World War. What was the primary purpose? To sink more German submarines or to get more ships safely across the Atlantic Ocean? An investigation of the efficiency of the anti-aircraft guns on British merchant vessels in the Mediterranean proved that the number of German airplanes shot down was very small, for which reason the efficiency of the guns might be questioned. The essential thing was, however, not the number of airplanes shot down but the number of sunken ships. For ships without anti-aircraft defense, the casualty figure was 25 percent; for vessels armed with anti-aircraft guns, only 10 percent.

In another example, an operations analyst in England was asked to try to improve the efficiency of the maintenance work on airplanes. He happened to be a biologist and thought of the problem in terms of the lifetime of a certain airplane. After a certain number of flying hours an airplane is overhauled and is then ready for flight again. The number of airplanes "ready for flight" was kept as high as possible. The objective was 75 percent (but 65 percent was more common). Flying crews presented no problem. The total number of airplanes

however, was limited. The biologist, proved that the criterion "airplanes ready for flight" was not the principal matter. The thing desired was "flying airplanes." (By preventing all flying one could soon obtain 100 percent of "airplanes ready for flight!") Therefore, they went to the opposite extreme. Every airplane that was "ready for flight" was actually flown. The percent of "airplanes ready for flight" decreased, but more airplanes were flown. The analysis resulted in doubling the striking force without a further increase in airplanes and ground personnel.

An example of a wrongly stated technical demand which can be mentioned is a requirement for a light anti-aircraft gun to protect a bridge, with the gun on the bridge. What one really wants is protection for the bridge, and it may be better protected with a heavier gun placed some miles away from the bridge.

PERSONNEL AND ORGANIZATION

To solve the above problems it is necessary to use personnel of very high quality with analytical ability and varied experience. At least three categories ought to be represented, namely, scientists, engineers and "users." These people ought to have a high general intelligence, a critical ability for observation, a good knowledge of scientific methods, the power to see problems from a broad point of view, an urge to experiment, and, last but not least, the ability to present results in an accessible way. Someone has said that an ideal operations research group should have Charles Darwin for the field work, Karl Pearson for the statistical work, and Bertrand Russell for the logical test of conclusions and the intellectual stimulus to the other members of the team.

System analysis is more a matter of attitude than a method that can be learned in a textbook. Without this attitude, even the most scientific statistical analysis may lead in the wrong direction. System analysis, like other operational research methods, is a staff function that ought to be placed near the top of an organization. There must not exist any restrictions on access to information due to security reasons, for example. The team should have freedom to do its work without pressure. Scientists cannot work effectively at the same speed that officers and engineers generally do.

THE IMPORTANCE OF SEEING THE WORK FROM A BROAD POINT OF VIEW

It is very important to see the problem as a whole when different elements are interconnected. The intention is to balance all the components so that the total efficiency will be as high as possible. A weapon is only a component in a larger military system. The problem of whether or not the World War II British bombers should have been equipped with defrosting gear when bombing Germany is an example. When the advantage of defrosting was weighed against diminished bomb load, it was found that the loss in lives and airplanes would be less without defrosting. Fewer flights would be needed to destroy a target, so flying personnel would be exposed to fewer risks.

QUANTITATIVE ANALYSIS

An analysis must not only be qualitative, but also quantitative. It shall not only produce adjectives! The famous English scientist, Lord Kelvin, once said: "When you can measure what you are speaking about and express it in numbers, you know something about it, but when you cannot measure it, when you cannot express it in numbers your knowledge is of a meagre and unsatisfactory kind." If system analysis is to be a scientific tool, one is obliged to have measurable and comparable values. For one needs numerical values in order to treat a

391368 O - 56 - 2

problem with statistical and mathematical methods. The most important elements must be sorted out and the immaterial ones thrown away. If there are too many basic elements the problem will be very difficult to analyze. It is therefore necessary to separate the primary and secondary factors. There is a common tendency for military authorities to construct a weapon that can do many different things. One must be sure that the weapon is not designed so that it has its maximum efficiency in an area of variables where the possibilities of success are small. What may happen if one specialist gets the upper hand is shown in the figure below.

Figure 1 - The specialist menace

As an example of a classification of demands on the construction of a field gun, the following can be mentioned:

Class I	Class II	Class III
1. Effect on the target	1. Rate of fire	1. Life span
2. Active range	2. Weight	2. Necessary production capacity (material and working hours)
3. Probability of hitting	3. Muzzle flame and smoke	
4. Fitness for field service		

A complete judgment is formed partly on fairly reliable technical data about the actual weapon and partly on other factors that are more difficult to obtain and which, in many cases, necessitate an essential simplification of the analysis. Therefore, it is often necessary for the estimation of different weapons in battle to define a limited number of "normal" battle situations. For every battle situation it is, as a rule, also necessary to assume far-reaching restrictions and definitions of the capabilities of one's own side as well as those of the opponent.

In this way one tries to define a technical battle situation so that it can be described mathematically. The difficulties of defining battle in a mathematical way can, to a certain extent, be compensated for by investigating different alternatives for the battle. In this connection, game theory may give valuable results. A battle between bombers and air defense can be described as a game where the attacking forces and the defending forces have different moves to make and each side tries to choose its moves so that it obtains optimal payoff even if the opponent is making an optimal move.

In "Naval Research Logistics Quarterly" for June 1954 is described a method for choosing from a class of weapon systems of a fixed cost that system which maximizes the expectation of military utility. The estimation of different factors is facilitated by the occurrence of operative constants. Thus, the number of mines per lost ship has been shown to be the same for German mines in the English Channel as for British mines in German waters and for American mines in Japanese waters. The reason for the occurrence of these constants is the human factor's dependence on the technical. Machines enforce certain rules that limit the individual variations and equalize the effects of such factors as fatigue, weather, etc.

THE CREATION OF A MATHEMATICAL MODEL

A mathematical model is created when the different alternatives for solving a problem have been examined. The technical world we live in is very complicated, but for practical reasons the working models must be relatively simple.

Mathematical models that describe a battle between two forces have been formulated by one of the pioneers in operations research, F. W. Lanchester, in Aircraft Warfare (London 1916). The most important of his laws states that the effective force is proportional to the effectiveness of the weapons multiplied by the square of the number of units engaged in the battle. The landing operation on Iwo Jima has been analyzed with Lanchester's theories and has given striking results in comparison with the actual events (JORSA 2:163, 1954). Lanchester's theories are particularly useful when a great number of participants are involved. But when only a few units are involved, as in sea or air battles, Markoff models have proven to be more suitable. In such a model it is possible to take into consideration many different factors. A battle at sea or a duel between two tanks or airplanes can be divided into different stages by deciding the transition probability between the various stages.

USING ELECTRONIC COMPUTING MACHINES OR CURVES

To solve the great number of optimality problems that usually are obtained in system analysis, electronic computing machines may be required. Sometimes, however, the use of simple curves proves effective. In Figure 2 is shown an example of a simple analysis of a two-motor civilian transport airplane (after a lecture of A. E. Raymond to the Royal Aeronautical Society on 10 September 1951). The demands are: maximum-speed at lowest height, 800 km/h; range, 2400 km; top height with one motor, 3800 m; take-off distance to 15-m height, 2000 m; and landing distance from 15-m height, 1350 m. An airplane inside the shaded area fulfills the desired conditions. If the requirements are set too high there will be no shaded area, and such a plane is impossible to construct.

THE PRESENTATION OF THE SOLUTION TO DECIDING AUTHORITIES

When the problem is solved it must be translated into intelligible language. All solutions must be put forward, even those that look absurd at a first glance. A solution should not be

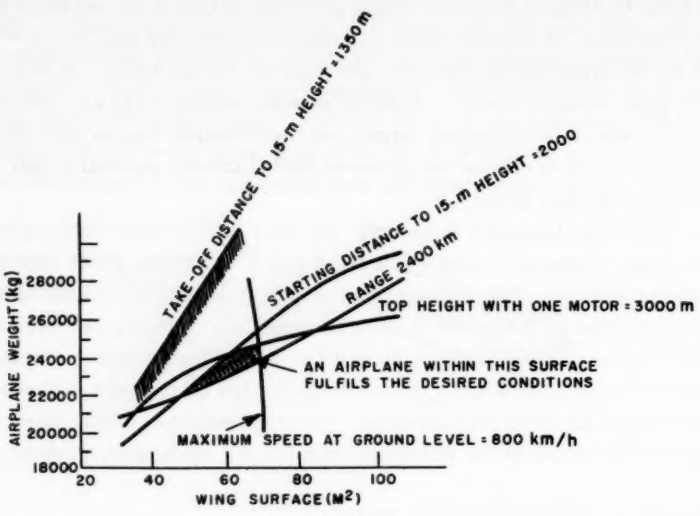


Figure 2 - Simple analysis of a two-motor civilian transport airplane

rejected only because it has been tried and previously found inappropriate. The circumstances may have changed.

The analysis should only give facts about the best way to reach a desired goal. The analysts have no responsibility for the decisions. How the results of the analysis shall be evaluated and judged is the problem of the deciding authority. Elementary mathematics is often sufficient to understand a problem, although higher mathematics may be required to solve it. Was it not Mark Twain who remarked that you don't have to be able to lay an egg in order to recognize it as rotten? The deciding authority, however, should have a certain amount of education in statistics and be familiar with the methods of scientific thinking. He must have a desire for objectivity. He must not be inclined to use only those results that fit into previously accepted theories and opinions and, with utter contempt for facts, throw away inconvenient results into the writing table drawer or into the waste basket!

COMMON SENSE

One can ask oneself if it is not easier to reach a better solution far quicker with common sense than with system analysis. What is meant by common sense? Think of two individuals with diametrically different opinions about the same thing. Each will surely call his opinion a result of common sense. Common sense is not always to be trusted. Common sense tells us that the earth is flat as a pancake. It also tells us that a bit of steel is solid. An X-ray photo shows, however, that it consists mainly of empty space.

He who does not entirely trust his judgment shows, thereby, good judgment! Even the soundest judgment must always be distrusted, for it is impossible for one person to collect all the elements needed to make it infallible. Error sources always sneak into all decisions; first, because they are never as reasonable as one tends to believe and always contain a certain amount of emotional thinking and, second, because they are founded on incomplete information. Pessimists consider common sense as rare as geniuses. A great number of opinions are fostered by defective statistical education which makes us pay too much attention to trifles. Mathematical and statistical methods, on the other hand, admit that data can "speak for themselves" without being affected by human errors. The power of mathematics lies in the fact that one is forced to stick to facts through which the number of wishful dreams is reduced. The advantage of an unprejudiced mathematical analysis is that it will also be impartial. Mathematics does not choose sides! It replaces judgment with figures. This does not mean that one shall not use common sense and good judgment. One must strive for a combination of mathematical analysis and common sense that can be called scientific or quantitative common sense. Common sense can be channeled so that it works with better effect.

One of the greatest advantages of system analysis is that one is forced into active thinking and must go to the bottom of the problems. It is quite common for human beings to try to find tricks to save themselves from thinking through their problems coherently and aimfully. To say that men are thinking beings is a poetical exaggeration. Human beings are "actually" very lazy and try, quite naturally, the most simple remedy which, however, is not always the best one. Thinking is troublesome and consumes much more energy than is generally believed. Despite the fact that the brain only represents 2 percent of the body weight, it consumes normally about 25 percent of the total energy! Generally, it is individuals with an instinctive unwillingness for active thinking and a dislike for thoroughness who "react" against every systematization of work.

Unfortunately, there are very few human beings who have what Socrates calls daimonion—the interior divine voice that warns when knowledge is insufficient for decision. Many executives make their decisions on the basis of old judgments and do not question whether or not these more or less traditional solutions really are the more effective. Human beings who prefer routine thinking and personal prejudices are not at all rare. Apparently self-evident things often need to be examined, and this is one of the definitions of a scientist: one who studies things that others regard as self-evident. A trained scientist is always hunting for error sources and tries to proceed only on clear rules and formulas. Through thinking in terms of numbers instead of words, subjective "factors" are avoided. An analysis of accessible data according to mathematical methods very often gives viewpoints and experiences that otherwise would have been overlooked. An undeveloped mind easily makes quick decisions and has very definite opinions. A scientist, on the other hand, is skeptical, expresses himself very carefully, and talks only about relative circumstances. Also, it is generally true that the less information one has the easier it is to make a decision. A deaf man has less trouble choosing a wife than a man with normal hearing, owing to the fact that the former has fewer aspects to take into consideration. But facility in making a decision is not the same thing as making the best decision! Knowledge is power, but it requires effort to get to the facts and not make only loose assumptions. Pascal says in his "Reflections" about mathematical analysis: "One never ought to give approval to other than proven truths, but most people are always disposed to believe not the things that rely on evidence but the things that attract them."

Even if a scientific analysis can't give exact answers and results, it does lead to an improvement in the procedure of thinking. One is obliged to examine all facts. The planning of the different steps in the analysis compels systematization. It is necessary to think in terms of components through which one is obliged to consider which information is required and what the next step will have to be. Since most people are bound to traditional trains of thought, it is useful to look at one's own problems anew through the application of experience from many sciences. Outside the scientific world there frequently is a powerful resistance to solving a problem in an objective and unprejudiced way. Many people will not accept a method that can seduce them into drawing conclusions which clash with favorite prejudices and intellectual hobby-horses.

Human sluggishness and spiritual friction are extraordinarily great. It is not the technical difficulties that offer the hardest resistance to the introduction of new methods, but the human beings that are governed by traditional ways of thinking. This is especially true for individuals that consider themselves "experts" in their fields. It is extraordinarily easy to be satisfied and say that so or so is absolutely the best method and that anything else is just nonsense. Many executives claim that they have no time to go to the bottom of a problem because they are overloaded with red tape and routine work. In the long run, however, they just get more to do if they decide in favor of a poorly analyzed construction, for this will require alternations, increased maintenance, and so on, and, above all, involve a lessened confidence in the responsible executive.

SUMMARY

An executive must today, more than ever, make conscious efforts to analyze available data. Because more complicated factors have to be taken into consideration and frequently 5 to 10 years elapse between the conception and production of a complicated weapon, an executive can't personally take time to grasp all details. He needs help in arranging the material to facilitate the making of decisions. And optimal effect of a series of decisions and actions is obtained in the safest way through a systematic analysis of the whole problem as well as its component parts. It is a wise man who has his afterthoughts first.

In Sweden, system analysis has been used by The Swedish Aeroplane Company (SAAB) to project new planes, and the favorable results obtained have caused a great interest in its continued use.

* * *

THE DYNAMICS OF OVERHAUL AND REPLENISHMENT SYSTEMS FOR LARGE EQUIPMENTS

Joseph Fennell and Seiki Oshiro[1]
The George Washington University
Logistics Research Project

1. INTRODUCTION

1.1 This paper is the result of preliminary investigation into the problem of scheduling aircraft purchases. During the early phases of the study it became clear that certain elementary questions which relate to this kind of scheduling could be handled rather easily by using techniques based upon a part of probability theory. Our only explicit intent here is to describe and exemplify certain methods which are useful in getting answers to questions like those cited in paragraph 1.3.

Because of the extreme generality of the problem underlying the topics actually treated (it is essentially the same one which might be expressed in such apparently diverse questions as "How many five year olds will have graduated from college by 1981?"; "Given a certain stake, how does the probability of breaking the house compare with that of going broke?"; or "Under a certain replacement policy how many electron tubes will be operative six weeks hence aboard some given ship?"), and by reason of the wide applicability of the tools used, it was decided to include certain observations, notably those of sections 4 and 9, which, largely for quantitative reasons, are not especially germane to the principal context.

The examples given below are necessarily limited in number and in the variety of their treatment. It is expected, however, that the discussion is sufficiently detailed as to indicate many of the generalizations and applications which might be made.

The matrix $P^{\#}$ introduced in the latter part of this paper defines a system sufficiently descriptive of actual practice to serve our present purposes. For reference and comparison the reader is referred to [3] in which a class of matrices based upon standard overhaul and retirement practice for Naval aircraft is defined. These matrices were equipped with entries previously derived by the Navy from operational data, and a set of five vectors much like the $\{\phi_k\}$, $\{\psi_k\}$ as defined below, were computed for the Navy for each type of plane for periods somewhat in excess of the statutory life spans.

1.2 Our reference throughout will be to aircraft, although the considerations apply equally well to the case of any relatively valuable equipment which is subject during its lifetime to a sequence of operating periods followed by overhauls.

1.3 A particular kind of ship is purchased and is destined to operate for a certain period of time, say N months. At the end of that time, barring washout and emergency removal for overhaul in the interim, an overhaul period of n months is begun. At the end of overhaul the ship is returned to operations. After a certain number of repetitions of this cycle the plane is retired. Given certain initial information such as "so many planes are in the i-th month of

[1]Work done under sponsorship of the Office of Naval Research. The authors wish to acknowledge their indebtedness to Captain S. R. Brown, Mr. C. E. Ellis, Captain W. E. Gentner, and Mr. A. E. Wence of OP-50 for the problem considered and for the stimulating discussions at that office which prompted the preparation of this material.

the first operational period, so many in the j-th month of the second overhaul, and so many new ships of the same kind will be introduced in the k-th month from now" one might inquire as to how many ships will be operational six months from now, how many will be in repair, how many will be due for retirement in nine months, and so forth. The methods described below show how answers may be obtained for a great variety of questions of this general nature.

1.4 Example I below is intended as an introduction to the notions and techniques which may be employed. It is not intended as a description of an actual working situation. It does suffice, however, to exhibit most of the ideas employed in the more elaborate model of paragraph 6.

1.5 The Concept of State. The idea of the states of a system is a very general one, with a wide range of applicability. For instance, a pair of dice is tossed, and we say that the system (the pair of dice) is in state n if the spots showing sum to n. We would then have a system with eleven states $S_2, S_3, \ldots S_{12}$. In what follows we shall be interested in the probability of a passage from one state to another, given that the system is in the former. In the language of the example, a gambler might want to know the probability of getting a 7 on the next roll, given that he has just tossed a 10. Here the system is in state 10 "at time 1," and he wants to know the probability that it will be in S_7 "at time 2," i.e., that he will next throw a 7.

Another place where the concept of state could be applied is given by the following example: The system is now a globe of heated gas. The "state" of the system is defined by counting the number of gas molecules in the upper hemisphere. We could just as well consider a system of states defined by listing the velocity and location of each molecule in the chamber. In general, the reasons for considering the system will dictate to a large extent just what kind of states are to be defined.

In the case at hand our states will be the more or less natural ones which might be used to describe the life of an aircraft, such as "the first month following delivery," "the second month of the first scheduled overhaul," and so forth.

1.6 Stationary Transition Probabilities. In paragraph 1.5 was mentioned the probability of getting a 7 following a throw of the dice which resulted in a 10. In the considerations which follow it will be necessary to refer to successive events of this kind by using expressions like "at time 0," "at time n" etc. In doing so we should like our statements to possess a certain kind of invariance with respect to time. More precisely, we shall stipulate that if the probability of the transition $S_{10} \to S_7$, given that the system is in S_{10} at time 0 is $P_{10,7}$, then the probability of the transition $S_{10} \to S_7$, given that the system is in S_{10} at time 11 (11 rolls later), is still $P_{10,7}$. In other words, the dice are not being changed, and all the participants in the game toss them in the same way. If we let Σ stand for system we can write the above requirement as follows:

$$\text{(1.6.1)} \qquad \Pr^0(S_{10} \to S_7 \mid \Sigma \text{ is in } S_{10}) = \Pr^{11}(S_{10} \to S_7 \mid \Sigma \text{ is in } S_{10}) = P_{10,7}\,.$$

The above probabilities all refer, of course, to "one step" (on the very next roll) transitions. It will be seen below that when (1.6.1) holds for the one-step case the same will be true for transitions over several time intervals.

A passage from peace to a state of war could cause an increase in the probability of passage from an operating state to washout. Simple changes in the technique in use here can be employed to handle our problem for the transition period.

1.7 The Markov Property. Several references have been made above to the state occupied by the system, mainly in the definitions of the conditional probabilities. It is quite conceivable that more of the historical data has a bearing upon the value of the transition probability. (For instance, it may become evident that information going beyond the state currently occupied by an aircraft has a real influence upon its behavior. Certain of these latter difficulties will be obviated automatically through the design of our models. Treatment of the more general case is quite within the scheme of things and involves only quantitative changes in method.) That is, we might be forced to consider questions of the form "given that Σ was in S_5, S_6, S_7 at times 0, 1, 2, respectively, what is the probability that it will be in S_9 at time 3?" Such questions can be converted into ones involving only the state occupied at present by the system. However, these considerations will not be important below. In all the cases considered we shall have,

(1.7.1) $\Pr(S_7 \to S_9 \mid \Sigma \text{ in } S_5, S_6, S_7 \text{ at times } 0, 1, 2) = \Pr(S_7 \to S_9 \mid \Sigma \text{ in } S_7 \text{ at time } 2)$.

Systems for which (1.7.1) holds in suitable form are termed Markov.

2. EXAMPLE I

2.1 An aircraft is scheduled to operate for six months following its purchase. At the end of that time it will receive an overhaul which lasts three months, following which it will reenter the system as a "new" ship. If it should washout at some time (it must washout from an operating state) it is reborn as a new plane after two months—that is, a new craft will take its place at that time. If it should require an unscheduled overhaul, this will be performed in the same manner as a scheduled one and will require the same amount of time. We shall describe this system by means of nine states $S_1, \ldots, S_9$, defined as follows:

S_1: The ship is in overhaul but has not completed the first 30 days there.

S_2: The ship has completed 30 but not 60 days of overhaul, or the ship has washed out but has not completed the 30th day of its reentry period.

S_3: The ship has completed 60 days overhaul but has not reentered the system as an operating craft, or the ship has completed the 30th day of the reentry period for washouts but has not reentered the system as an operating craft.

S_4: The ship has entered the system new or has reentered from S_3 but has not completed the 30th day of operations.

S_5, S_6, S_7, S_8: The ship has completed 30, 60, 90, 120 but not 60, 90, 120, 150 days operations, respectively.

S_9: The ship has completed 150 days operations but has not entered its scheduled overhaul and has not washed out.

2.2 It is implicit in the above that the events of entry into the system, washout, and emergency overhaul can occur only on the 30th, 60th, etc. day counted from the starting point. Actually, this is little in the way of a practical limitation, since we could just as well have considered the system day by day and enlarged our system of states about thirtyfold. Also, it is unnecessary for us to carry along the expressions, 30th day, 60th day, and so on. We can just as well use the phrases time 1, time 2, etc., in all future references of this kind. Figure 1 is a schematic representation of the possible paths open to a ship in the system just described.

2.3 From Figure 1 it would appear that certain of the system events are sure. A ship in state 1 must pass to S_2 and from S_2 to S_3. In some of the states, for instance state 6, the

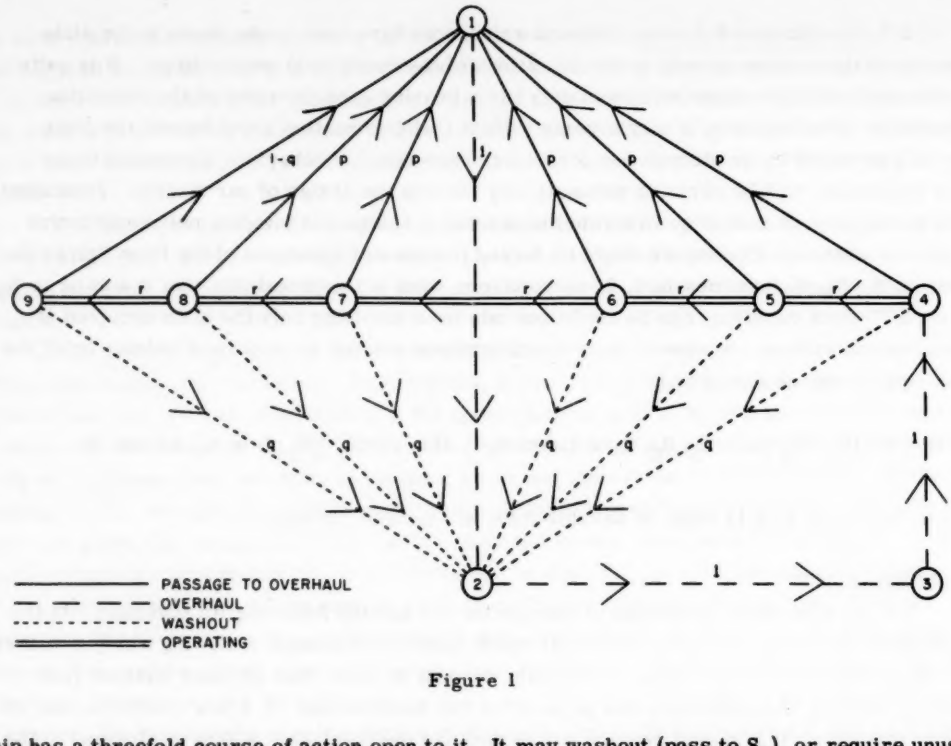


Figure 1

ship has a threefold course of action open to it. It may washout (pass to S_2), or require unusual overhaul (pass to S_1), or simply survive and proceed to the next operating state (pass to S_7). To complete our model we shall assign probabilities to the events of washing out and of requiring nonregular overhaul. Let these be q and p, respectively. Using Figure 1 and the numbers p, q, it is now possible to describe our example completely by means of a matrix P. Some of the advantages in doing this will appear below.

P:

	1	2	3	4	5	6	7	8	9
1		1							
2			1						
3				1					
4	p	q			r				
5	p	q				r			
6	p	q					r		
7	p	q						r	
8	p	q							r
9	1-q	q							

The indices on top and to the left of P refer to the various states, 3 to S_3, and so on. The (i, j) entry of P is the conditional probability that: given that a ship is in state i at time n it will be in state j at time n + 1. Evidently, since by definition an aircraft in any given state must pass to another of the system, the entries in each row must sum to 1. In particular, r,

the probability of survival from one operating state to the next, must equal 1-p-q. Entries missing in P are zero. One may check P against A, noting that a particular entry (i, j) is non-zero if, and only if, an arrow leads from i to j in Figure 1.

2.4 In the above the probability of a passage from S_4 to S_5 has been set equal to that for a passage from S_5 to S_6. If we should want to consider a model wherein the chances of washout and unusual overhaul increase (or decrease) as the end of an operational tour is neared, corresponding changes could be made in P. The diagonal of r's would be changed to r_1, r_2, etc., and suitable adjustments made in the p's, q's.

2.5 The model just described evidently possesses the features of no permanent retirement for aircraft and of automatic replenishment with a fixed lead time for replacement of washouts. Before passing to systems more descriptive of reality we shall discuss certain properties of the probabilities P_{ij} and of matrices like P.

3. POWERS OF P AND HIGHER TRANSITION PROBABILITIES

3.1 The matrix P describes a Markov chain.[2] Certain properties of such matrices will be useful below, and they will now be described.

3.2 Suppose we denote the (i, j) entry of P by p_{ij} so that $p_{11} = 0$, $p_{12} = 1$, $p_{41} = p$, etc. Then, as already noted, p_{ij} is the conditional probability that: the system will be in state j at time n+1, given that it is in state i at time n. One could inquire concerning the probability of a passage from i to k in two time periods. Suppose Σ, the system, is in S_i at time n. Then, in order to be in S_k at time n+2, Σ must occupy an intermediate state S_j at time n+1, where j can be any of the indices 1, 2, 3, . . . , 9. Using the usual formulae for manipulating conditional probabilities and the Markov property noted in paragraph 1.7, the probability of the passage $S_i \to S_j \to S_k$ is found to be equal to $p_{ij} \cdot p_{jk}$. Since j can be any of nine indices, the probability of the two-step transition $S_i \to S_k$ is obtained by summing the expressions $p_{ij} \cdot p_{jk}$ over j. We have then,

$$p^2_{ik} = \sum_{j=1}^{9} p_{ij} \cdot p_{jk} . \tag{3.2.1}$$

The superscript 2 on p^2_{ik} stands for the time lapse and p^2_{ik} is the conditional probability that: Σ will be in S_k at time n+2, given that it was in S_i at time n.

3.3 The above expression for p^2_{ik} will be recognized as that for the (i, k) entry of the matrix product of P with itself. In other words, the two-step transition probabilities for the system Σ are just the entries of $P \times P$ computed by the rule (3.2.1). Thus p^2_{41} would be equal to $p \cdot 0 + q \cdot 0 + 0 \cdot 0 + 0 \cdot p + r \cdot p + 0 \cdot p + 0 \cdot p + 0 \cdot p + 0 \cdot (1-q) = rp$. This corresponds to the situation depicted in Figure 1, wherein the only way to get from S_4 to S_1 in two steps is by going to S_5 (probability r) and from there to S_1 (probability p). In this connection it should be noted that a passage like $S_i \to S_k \to S_k$ is a perfectly valid method for getting from S_i to S_k in two steps. It cannot actually happen in the case at hand, since all our diagonal elements p_{kk} are zero.

3.4 It can be shown rather easily that the transition probabilities for an arbitrary number of steps are obtainable in the same way as the p^2_{ij}. In general p^n_{ik} (the conditional probability that Σ will be in S_k at time n, given that Σ was in S_i at time 0) is given by the formula (3.4.1).

---

[2] A beautiful and readable account of certain kinds of Markov chains is given in [1].

$$(3.4.1) \qquad p^{n}_{ik} = \sum_{j=1}^{9} p^{n-1}_{ij} \cdot p_{jk} = \sum_{j=1}^{9} p^{r}_{ij} \cdot p^{s}_{jk}$$

where $r + s = n$, $r, s \geq 0$, $p^{0}_{ij} = 0$ unless $i = j$ and $p^{0}_{jj} = 1$, and where p^{r}_{ij} (p^{s}_{jk}) is the (i, j) (the (j, k)) element from P^{r}, (P^{s}). P^{r} (r any positive integer) stands for the matrix product,

$$\overbrace{P \times P \times \ldots \times P}^{\text{r terms}}.$$

3.5 It can be readily demonstrated that the row sums of P^{n}, $n > 1$, are again unity. For the case $n=2$ we have

$$\sum_{k=1}^{9} p^{2}_{ik} = \sum_{k=1}^{9} \left(\sum_{j=1}^{9} p_{ij} \cdot p_{jk} \right)$$

$$(3.5.1) \qquad = \sum_{j=1}^{9} p_{ij} \sum_{k=1}^{9} p_{jk}$$

$$= \sum_{j=1}^{9} p_{ij} = 1 .$$

The cases $n > 2$ can be handled in the same way, using the inductive hypothesis: $\sum_{k=1}^{9} p^{r}_{ik} = 1$, $r < n$; and the formulae (3.4.1).

3.6 It is of some minor interest to note the rather simple form which the p^{n+1}_{ij} have when expressed in terms of the p^{k}_{ij}, $k \leq n$. We have directly from P that

$$(3.6.1) \qquad p^{n+1}_{ij} = \begin{cases} p^{n}_{i+1,\, j} & \text{, if } 1 \leq i \leq 3 \\ p \cdot p^{n}_{1j} + qp^{n}_{2j} + rp^{n}_{i+1,\, j} & \text{, if } 4 \leq i \leq 8 \\ (1-q)\, p^{n}_{1j} + q \cdot p^{n}_{2j} & \text{, if } i = 9 \end{cases}$$

4. SOME FURTHER REMARKS CONCERNING P

4.1 There is an additional feature of P having an intrinsic significance somewhat removed from any contextual interest. Its statement in paragraph 4.2 does, however, call attention to the rather important question of system behavior over long periods of time. It serves also to exhibit the correspondence between the mathematics of the model and what might reasonably be inferred or deduced experimentally about this kind of system when it is allowed to operate for several time periods.

The passage of Σ to what might be called its steady state is illustrated in Table 1 and in the bar charts of paragraph 4.6.

4.2 In the case of the matrix P a certain theorem[3] states that $\lim_{n\to\infty} p_{ij}^{n}$ exists for each (i, j) and that these limits are row independent quantities u_j satisfying the system of linear equations (4.2.1).

$$u_j = \sum_{i=1}^{9} p_{ij} u_i \; ; \quad \sum_{j=1}^{9} u_j = 1 \, . \tag{4.2.1}$$

This system may be solved for $u_2, u_3, \ldots, u_9$ in terms of u_1. Finally, the requirement $\sum_{j=1}^{9} u_j = 1$ is used to determine u_1. The solutions are as follows:

$$u_1 = \frac{1 - \dfrac{q(1 - r^6)}{1 - r}}{3 + \dfrac{(1 - q)(1 - r^6)}{1 - r}}$$

$$u_2 = \frac{1}{3 + \dfrac{(1 - q)(1 - r^6)}{1 - r}} \tag{4.2.2}$$

$$u_3 = u_4 = u_2$$

$$u_{j+4} = r^j u_4 \, , \quad 1 \le j \le 5 \, .$$

4.3 The kind of system described in Example I can be enlarged easily to one in which the length of the operational tour is $\ell + 1$, overhauls require k time periods, and washouts are replaced in $m < k$ time periods. The solutions corresponding to (4.2.2) for the more general case are given by (4.3.1).

$$\begin{aligned}
u_1 &= A/\left[(k - m)A + m + (1 + r + \ldots + r^{\ell})\right] \\
u_i &= u_1 \, , \quad 2 \le i \le k - m \\
u_i &= u_1/A \, , \quad k - m + 1 \le i \le k + 1 \\
u_{k+1+i} &= r^i u_1/A \, , \quad 1 \le i \le \ell
\end{aligned} \tag{4.3.1}$$

where $A = p(1 + r + r^2 + \ldots + r^{\ell-1}) + r^{\ell}(1 - q)$.

4.4 The two preceding paragraphs mean simply this: That a system operated like that of Example I can be expected, following a certain initial time lapse, to have, from then on, $u_1 \cdot N$ aircraft in S_1, $u_2 \cdot N$ in state 2, and so on, where N stands for some initial number of

[3] See again [1], in particular, Chapter 15.

planes inserted into any state of the system and where replenishment of washed out ships is accomplished in the manner called for by the structure of P, or, in the case of paragraph 4.3, automatically after m time periods.

4.5 The powers P^n were computed for $n < 1022$, using $r = .97$, $P = .02$, $q = .01$. Table 1 lists certain of the results, to four decimal places. The computed P_{ij}^{1022} differed from the u_j, obtained by using (4.2.2), by about 10^{-8}.

TABLE 1

n	P_{11}^n	P_{12}^n	P_{13}^n	P_{14}^n	P_{15}^n	P_{16}^n	P_{17}^n	P_{18}^n	P_{19}^n
0	1	0	0	0	0	0	0	0	0
5	.0194	.0297	.0100	.0000	.0000	.9409	.0000	.0000	.0000
10	.0019	.8524	.0278	.0277	.0272	.0271	.0271	.0089	.0000
15	.0410	.0514	.0537	.0210	.0037	.0029	.7780	.0246	.0238
20	.0075	.0082	.7290	.0489	.0473	.0466	.0469	.0475	.0180
25	.0566	.0660	.0691	.0728	.0312	.0094	.0075	.6454	.0420
50	.0936	.1087	.0626	.0349	.0290	.4407	.0791	.0760	.0754
100	.0662	.2317	.1111	.1090	.1074	.1108	.1166	.0845	.0628
250	.1102	.1164	.1177	.1204	.1141	.1079	.1031	.1091	.1010
500	.1109	.1175	.1175	.1175	.1139	.1106	.1072	.1040	.1009
1000	.1109	.1175	.1175	.1175	.1140	.1105	.1072	.1040	.1009

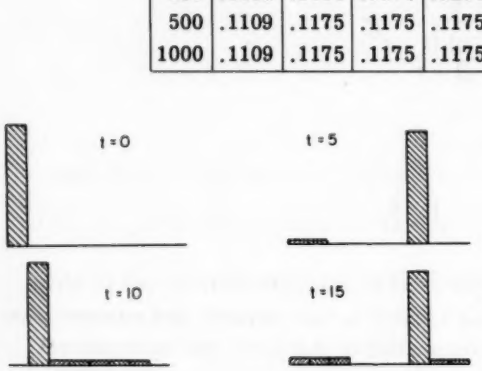


Figure 2

4.6 In the charts of Figure 2 the bar heights are proportional to those entries in Table 1, with the same time index t. They depict the distribution of airfraft in Σ following the insertion of a block of N planes into S at t=0. The height of the jth bar in the diagram for t=k is the expected fraction of the N aircraft in S_j at time k. A description of some of the rationale behind the use of the term "expected" follows in Section 5.

5. FURTHER REMARKS CONCERNING (Σ, P)

5.1 It will be convenient in this section to continue to refer to the system (Σ, P) of Example I, although the considerations carry over almost word for word to any of the systems being studied, and in particular to the case of Example II.

5.2 Suppose that one tossed a fair coin one hundred times. There is some intuitive feeling that 50 qualifies for the title of "expected number of heads." This terminology is traditional in a more formal sense. In the case of repeated trials of the same

kind where each result is one of two possible ones, say A and B, the expected number of A's is defined to be the quantity E(A) where,

$$E(A) = N \cdot p(A), \tag{5.2.1}$$

and where p(A) stands for the probability that a single trial will result in A and N for the number of trials. For a "fair" coin we have $p(H) = \frac{1}{2}$, and the figure 50 results from the computation (5.2.1) with N = 100.

5.3 In the case of the matrix P we have, by definition, that the probability that a ship in S_i at time t=0 will be in S_j at time n is P_{ij}^n. Picking one particular state, say S_1, we perform the "trial" of watching the ship until time n. Then either it is in S_1 or not in S_1, i.e., in $S_2, S_3, \ldots, S_9$. The probability of the former event is P_{i1}^n, while that of the latter is $1 - p_{i1}^n$, or, by (3.5.1), $p_{i2}^n + p_{i3}^n + \ldots + p_{i9}^n$. Starting with N aircraft in S_i at time 0, we can regard their careers until time n as repeated and independent (they presumably do not crash into one another too often) trials of the same kind. This is precisely the situation described in paragraph 5.2, and we call $N_1 = N \cdot p_{i1}^n$ the expected number of planes in S_1 at time n.

The same considerations apply to any separation of the states S_j into two disjoint classes. For instance, if we start with N planes in S_1 at time 0 and wish to compute the expected number in operation at time n, we have for the case of (Σ, P) that:

The probability that a plane in S_1 at time 0 is operating at time n is $p_{14}^n + p_{15}^n + p_{16}^n + p_{17}^n + p_{18}^n + p_{19}^n$, since $S_4, S_5, \ldots, S_9$ are the totality of operational states.

Similarly,

The probability that a plane in S_1 at time 0 is not operating (is in overhaul) at time n is $p_{11}^n + p_{12}^n + p_{13}^n$, since S_1, S_2, S_3 are the only non-operational states for Σ.

To take an example: Let N = 1000; then from Table 1 we have that:

The expected number of planes in S_1 at time 5 is 19
The expected number of planes in S_3 at time 10 is 28
The expected number of planes in S_4 at time 15 is 21
The expected number of planes in S_6 at time 25 is 9;

in operations at time 20 is 48.9 + 47.3 + 46.6 + 46.9 + 47.5 + 18 = 255; in overhaul at time 50 is 93.6 + 108.7 + 62.6 = 265.

5.4 The last two kinds of probabilities mentioned above are of particular importance in that they are basically related to aircraft build-up and abandonment-of-type programs as well as to questions concerning repair overload. We propose, therefore, to consider them in more detail and to illustrate their usefulness in connection with questions of aircraft programming.

5.5 It will be convenient to adopt certain simpler notation for the partial row sums used above. We shall also agree that all aircraft coming into the system are inserted into S_4, the first operational state. ϕ_k, ψ_k, the probability that a plane in S_4 at time 0 is operational, in overhaul, respectively, is defined as follows:

$$\phi_k = \sum_{j=4}^{9} p_{4j}^k \quad ; \quad k = 0, 1, 2, \ldots \tag{5.5.1}$$

$$\psi_k = \sum_{j=1}^{3} p_{4j}^k \quad ; \quad k = 0, 1, 2, \ldots . \tag{5.5.2}$$

Suppose now that a certain number N of new planes are inserted into S_4 of Σ at time 0. That part of the future operational strength of Σ which is due to this input can be read off for each time t=0,1,2,3,..., from a graph of the quantities $\phi_j \cdot N$ like that of Figure 3.

5.6 The stability of operational strength at a point greater than zero evidenced from about t=13 on is due to the automatic replenishment feature built into the model Σ. As we shall see below, the long-run behavior of other systems will depend upon the over-all nature of the set of inputs. Roughly speaking, input sequences which oscillate rather narrowly about some fixed level will produce stable operational levels if the operation of this system is continued for some time.

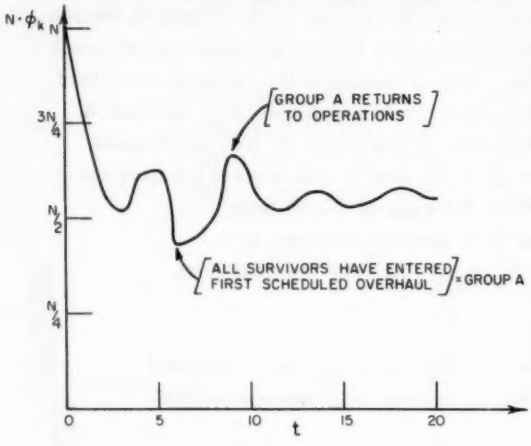


Figure 3

5.7 We consider now the future operational strengths of Σ when, starting with an empty system, N planes are inserted at t=0, and say N/2 at time 1. These figures can be read off the resulting graph when the two broken curves shown in Figure 4 are added. The result of the addition is shown in the unbroken curve of that figure.

5.8 Evidently, from the foregoing, one must add certain multiples of the "fundamental" curve shown in Figure 3 in order to obtain a given program of operational strength, and prior to the addition translate these multiples suitably to correspond to the time of the input represented. We next show one method whereby this may be accomplished for the system Σ.

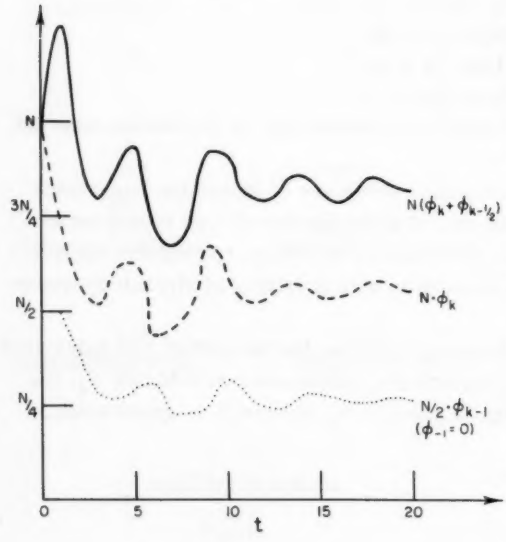


Figure 4

5.9 We start with an empty system Σ and imagine that N_0 planes are inserted at time 0, N_1 at time 1, and so on. Then from the definition of the ϕ_j we have:

The expected number of operational craft at t = 0 is $N_0 \cdot \phi_0$

The expected number of operational craft at t = 1 is $N_1 \cdot \phi_0 + N_0 \cdot \phi_1$

The expected number of operational craft at t = 2 is $N_2 \cdot \phi_0 + N_1 \cdot \phi_1 + N_0 \phi_2$

The expected number of operational craft at t = 3 is $N_3 \phi_0 + N_2 \phi_1 + N_1 \phi_2 + N_0 \phi_3$ etc.

In other words, we get the expected number of aircraft in operations at time k by multiplying N_k (those just inserted) by the probability that they will be in operations 0 time hence, N_{k-1}, the number of those inserted one time period ago, by ϕ_1, the probability of passage to an operational state in one step, and so forth, and by then adding these quantities.

Suppose some program is set down by specifying operational strengths desired at times 0, 1, 2, etc. Let G_0 be the number of operational aircraft desired at time 0, G_1 the number required at time 1, etc. One inquires as to the inputs $N_0, N_1, \ldots, N_k$ needed to produce the G_j. From the above this would appear to be very similar to the familiar process of adding up suitable multiples of cos nx, sin nx to obtain some predetermined curve. The functions graphed in Figures 3 and 4 correspond to cos nx, sin nx and the numbers N_j to the Fourier coefficients of cos, sin.

From the first paragraph of 5.9 it may be seen that in order to have G_0 operating planes at time 0 we must have $N_0 \phi_0 = G_0$; to have G_1 at time 1, $N_0 \phi_1 + N_1 \phi_0$ must equal G_1; and so on. Proceeding in this way we get a triangular system of linear equations which may be solved for the N_j in terms of the ϕ_k and G_i. Explicitly, the system becomes,

$$N_0 \phi_0 = G_0$$

$$N_1 \phi_0 + N_0 \phi_1 = G_1$$

$$N_2 \phi_0 + N_1 \phi_1 + N_0 \phi_2 = G_2$$

$$N_3 \phi_0 + N_2 \phi_1 + N_1 \phi_2 + N_0 \phi_3 = G_3$$

etc.

These equations can be solved readily for the N_j. In fact, since $\phi_0 = 1$, we get for solutions

$$N_0 = G_0$$

$$N_1 = G_1 - N_0 \phi_1$$

(5.9.1)

$$N_2 = G_2 - N_1 \phi_1 - N_0 \phi_2$$

$$N_3 = G_3 - N_2 \phi_1 - N_1 \phi_2 - N_0 \phi_3$$

etc.,

where N_1 is computed by substituting G_0 for N_0 in the second equation, N_2 by substituting G_0 for N_0 and $G_1 - G_0\phi_1$ for N_1 in the third, etc.

We now consider an example of this process. Our system is Σ, and we set

$$G_0 = 10$$

$$G_1 = 20$$

$$G_2 = 30$$

391368 O - 56 - 3

$$G_3 = 40$$
$$G_4 = 50$$
$$G_5 = 60$$
$$G_6 = 70$$
$$G_7 = 80$$
$$G_8 = 90$$
$$G_j = 100 \ , \ j \geq 9 .$$

The equations (5.9.1) become

$$N_0 = 10$$
$$N_1 = 20 - 10\phi_1$$
$$N_2 = 30 - \phi_1 (20 - 10\phi_1) - 10\phi_2$$
$$N_3 = 40 - \phi_1 [30 - \phi_1 (20 - 10\phi_1) - 10\phi_2] - \phi_2 (20 - 10\phi_1) - 10\phi_3$$

etc.

Solving these is not as cumbersome as it might appear. Using a desk computer one of the authors, in about three hours, obtained N_k for all k up to about 40. Part of the results are given below, and the resulting operational strengths are graphed in comparison with the G_j in Figure 5. The set of required inputs for the above program $\{G_j\}$ was found to be as follows:

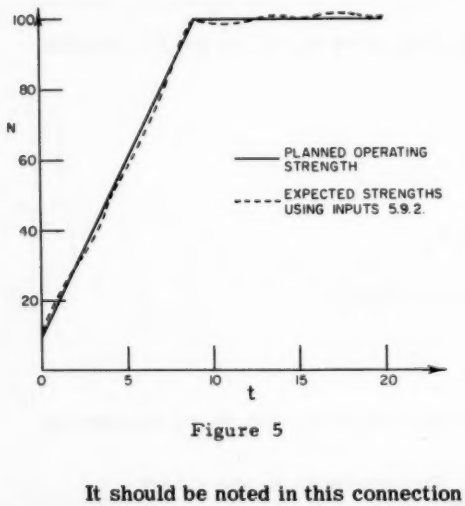


Figure 5

(5.9.2)

$N_0 = 10$ $N_9 = 18$

$N_1 = 13$ $N_{10} = 8$

$N_2 = 15$ $N_{11} = 5$

$N_3 = 16$ $N_{12} = 3$

$N_4 = 17$ $N_{13} = 2$

$N_5 = 16$ $N_{14} = 2$

$N_6 = 18$ $N_{15} = 2$

$N_7 = 19$ $N_j = 0 \ , \ j \geq 16 .$

$N_8 = 19$

It should be noted in this connection that the above N_j are not exact solutions to the system (5.9.1) because of the need for integer-valued N_j; and because we prefer not to admit negative inputs, correct solutions such as $N_j = -.5634$ must be replaced by $N_j = 0$. The use of negative inputs would change only the computational procedure. No technical difficulties arise.

5.10 In other words, to build up to and maintain the strength given by the above G_j one must insert about 183 planes into Σ. It should be noted that we cannot consider abandonment

programs in connection with Σ, since this system possesses no outlets. The figure of 183 could have been rapidly estimated for this system by computing 100/.5516, where 100 is the required "steady-state" strength and .5516 is $\sum_{j=4}^{9} u_j$, the sum of the steady-state probabilities for the operational states of Σ, as defined in (4.3.1).

5.11 The linearity of $\{N_j\}$ considered as a function of $\{G_j\}$ should be emphasized here. If one should desire a program of operational strength equal in each case to twice the G_j listed in paragraph 5.9, the corresponding N_j should be just twice those listed in (5.9.2). More generally, if the inputs $\{N_j\}$ produce an operational program $\{G_n\}$ and the inputs $\{M_j\}$ a program $\{H_n\}$, then the inputs $\{aN_1 + bM_1, aN_2 + bM_2, \ldots, aN_k + bM_k,\} = \{aN_j + bM_j\}$ will produce the operational strengths $\{aG_1 + bH_1, aG_2 + bH_2, \ldots, aG_n + bH_n, \ldots\} = \{aG_n + bH_n\}$, where a, b are any real numbers.

More compactly, we could write: if $N \longrightarrow G$ and if $M \longrightarrow H$, then $aN + bM \longrightarrow aG + bH$, where "$\longrightarrow$" can be read "produces."

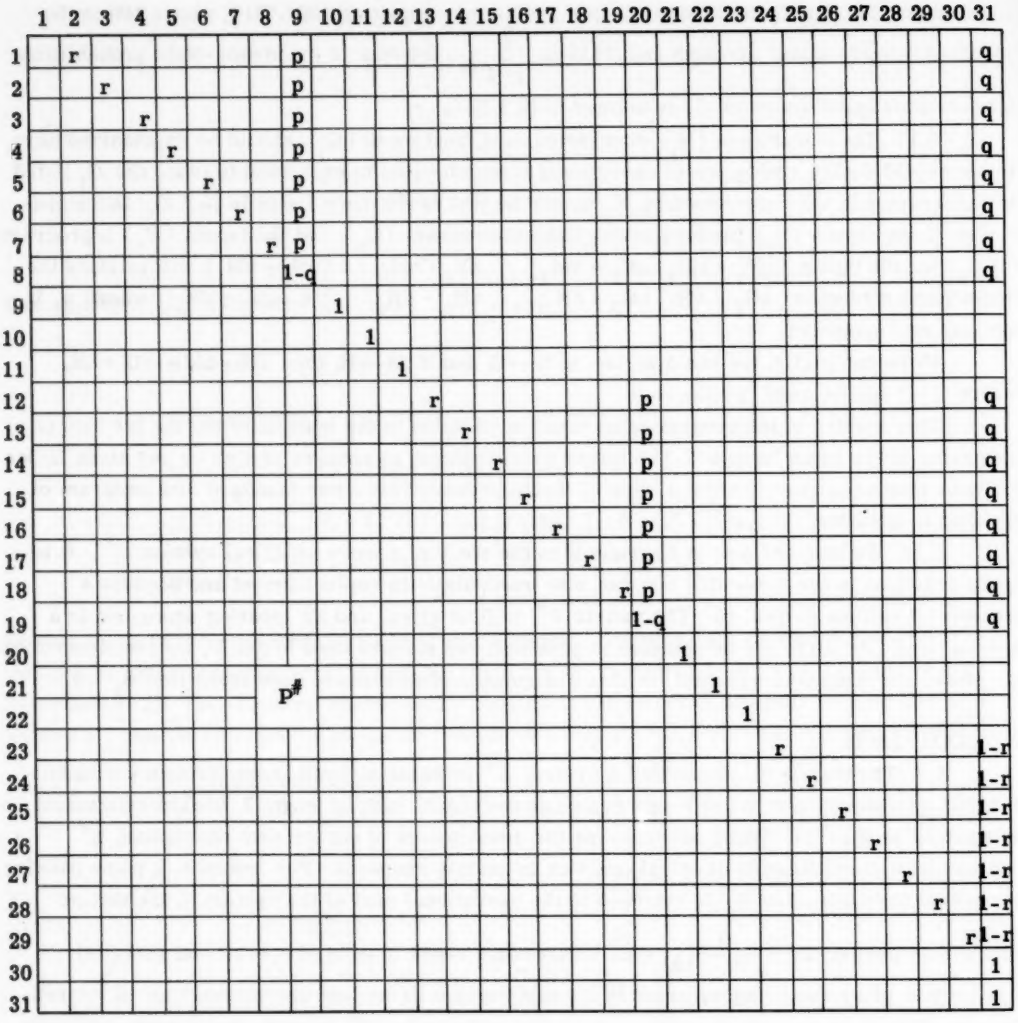
This feature of the programming could be of value in the feasibility testing for size of programs of the same "shape." The inputs corresponding to changes of size by any given factor could be computed readily from a table of inputs prepared for some standard size program of the kind in question.

5.12 We proceed now in Example II to the study of a more practical system $\Sigma^{\#}$. It is more practical in the sense that one can now manipulate its replenishment and because a retirement feature is built in. The matrix $P^{\#}$ is first given, and its meaning described in a general way. Its form can be obtained in precisely the manner used to get P, i.e., by describing the states one wishes to use, and by then assigning probabilities to each transition $S_i \longrightarrow S_j$.

6. EXAMPLE II

6.1 The matrix $P^{\#}$ describes a system $\Sigma^{\#}$ in which aircraft move through a schedule of three operational periods of length 8 and two overhauls each of length 3, toward retirement or washout in state 31. While adequate for the development of our present discussion, $\Sigma^{\#}$ represents a simplification of actual practice in certain respects. For instance, a plane passing to unusual overhaul from S_1 is returned to the operational part of the system in the second operational period, $S_1 \overset{p}{\rightarrow} S_9 \overset{1}{\rightarrow} S_{12}$, thus losing some seven months of operational potential. Similarly, a plane requiring overhaul in the early stages of its last operational tour is retired. These difficulties can be very simply avoided by redesigning $P^{\#}$ to represent a more elaborate system.

It should be noted that $P^{\#}$ has a "sink," i.e., state 31, into which retired and washed out aircraft proceed and from which there is no return ($P_{31,31} = 1$). Unlike Σ, $\Sigma^{\#}$ is not self-replenishing.



At time 31 all aircraft in $\Sigma^{\#}$ at time 0 will have been retired. This feature is reflected in the 31st power of $P^{\#}$, which consists of a column of 1's in the 31st column. Thus any continued operational strength for $\Sigma^{\#}$ can be obtained only through the continued insertion of aircraft supplied by industry into some state S_j, $j \leq 30$.

6.2 For reasons of space we shall not reproduce here the powers of $P^{\#}$, which involve about 30,000 lengthy decimals. We shall give in detail, however, lists of all the more important parameters computed from those powers.

6.3 We first consider a simple program of aircraft operational strengths G_j for the system $\Sigma^{\#}$. The G_j stand for the desired operating strengths at times j, and we list them for Program (IIa):

(6.3.1)

$$G_0 = 10 \qquad G_{51} = 90$$
$$G_1 = 20 \qquad G_{52} = 80$$
$$G_2 = 30 \qquad G_{53} = 70$$
$$G_3 = 40 \qquad G_{54} = 60$$
$$G_4 = 50 \qquad G_{55} = 50$$
$$G_5 = 60 \qquad G_{56} = 40$$
$$G_6 = 70 \qquad G_{57} = 30$$
$$G_7 = 80 \qquad G_{58} = 20$$
$$G_8 = 90 \qquad G_{59} = 10$$
$$G_j = 100, \ 9 \le j \le 50 \qquad G_i = .0, \ 60 \le i \le 70 .$$

6.4 $\Sigma^{\#}$ will be regarded as empty prior to time 0, and the required inputs N_j (we have decided arbitrarily to insert planes only into S_1) will be computed as in (5.9.1), where the ϕ_j are now defined as follows:

(6.4.1) $$\phi_0 = 1$$

(6.4.2) $$\phi_n = \sum_{j=1}^{8} p_{1j}^n + \sum_{j=12}^{19} p_{1j}^n + \sum_{j=23}^{30} p_{1j}^n \quad 1 \le n \le 30$$

(6.4.3) $$\phi_n = 0 \quad 31 \le n .$$

6.5 It should be noted that if we agree that $(P\#)^0$ = a diagonal matrix of 1's, and if we take into account the fact that p_{ij}^n, $j \le 30$, is 0 for all $n \ge 31$ the definitions (6.4.1, 2, 3) can be written more compactly in the form,

(6.5.1) $$\phi_n = \sum_{j=1}^{8} p_{1j}^n + \sum_{j=12}^{19} p_{1j}^n + \sum_{j=23}^{30} p_{1j}^n , \quad 0 \le n .$$

That is, the ϕ_n are, as in the case of Σ, simply the probabilities that a plane in S_1 at time 0 will be operating at time n, where $S_1, \ldots, S_8$, $S_{12}, \ldots, S_{19}$, and $S_{23}, \ldots, S_{30}$ are the totality of operating states for $\Sigma^{\#}$.

The computation of the ϕ_n (and that of the other quantities involved) was based upon the following values, $p = .10$, $q = .05$, $r = .85$.

The ϕ_n were found to be as follows:

(6.5.2)

$$\begin{aligned}
\phi_0 &= 1.0000 & \phi_{16} &= .3707\\
\phi_1 &= .8500 & \phi_{17} &= .3516\\
\phi_2 &= .7225 & \phi_{18} &= .3260\\
\phi_3 &= .6141 & \phi_{19} &= .2226\\
\phi_4 &= .6220 & \phi_{20} &= .2102\\
\phi_5 &= .6137 & \phi_{21} &= .1947\\
\phi_6 &= .5939 & \phi_{22} &= .2510\\
\phi_7 &= .5662 & \phi_{23} &= .1915\\
\phi_8 &= .2710 & \phi_{24} &= .1450\\
\phi_9 &= .2917 & \phi_{25} &= .1087\\
\phi_{10} &= .3073 & \phi_{26} &= .0806\\
\phi_{11} &= .5903 & \phi_{27} &= .0590\\
\phi_{12} &= .5006 & \phi_{28} &= .0424\\
\phi_{13} &= .4290 & \phi_{29} &= .0297\\
\phi_{14} &= .3714 & \phi_{30} &= .0000\\
\phi_{15} &= .3791 & \phi_{31} &= .0000
\end{aligned}$$

6.6 Solving the equations

$$N_0 \phi_0 = G_0$$

$$N_0 \phi_1 + N_1 \phi_0 = G_1$$

. . . . . .

as in section 5, the required inputs were found to be as follows:

(6.6.1)

j	N_j	j	N_j	j	N_j	j	N_j
0	10	13	7	26	9	39	9
1	12	14	6	27	8	40	9
2	13	15	6	28	10	41	8
3	14	16	8	29	10	42	8
4	15	17	7	30	10	43	9
5	16	18	5	31	11	44	8
6	16	19	4	32	10	45	8
7	16	20	4	33	11	46	8
8	20	21	7	34	10	47	8
9	21	22	7	35	10	48	8
10	11	23	8	36	10	49	9
11	8	24	8	37	9	50	8
12	7	25	9	38	9	51	9

With $N_j = 0$, $j > 51$. For precisely the reasons cited in paragraph 5.9, it is of interest to compare the quantities $\sum_{j=0}^{k} \phi_j N_{k-j} = \tilde{G}_k$ with the desired G_k as given in (6.3.1), and where the N_{k-j} are taken from (6.6.1). This is done graphically in Figure 6. In the same chart the load upon the overhaul subsystem for $\Sigma^{\#}$ also is given. This was obtained by computing, as in (5.5.2), the quantities ψ_n which bear the same relation to overhaul states that the ϕ_n do to the operational portion of $\Sigma^{\#}$.

(6.6.2) $$\psi_n = \sum_{j=9}^{11} p_{1j}^n + \sum_{j=20}^{22} p_{1j}^n .$$

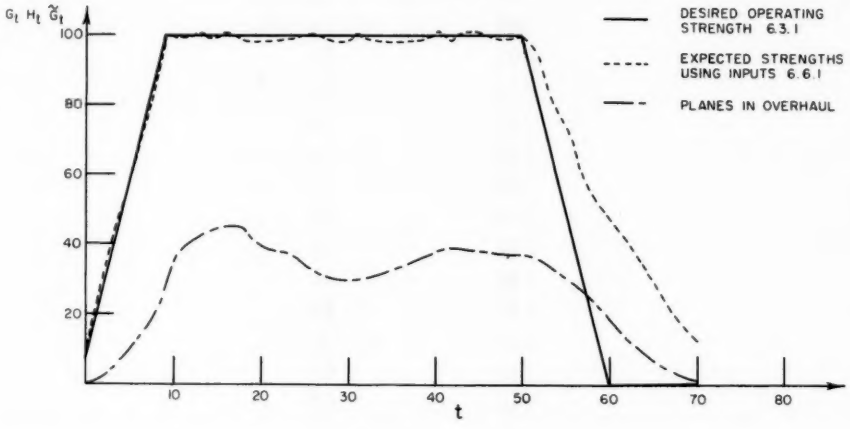


Figure 6

The load upon the overhaul system at time k is then given by,

(6.6.3) $$H_k = \sum_{j=0}^{k} \psi_j N_{k-j} .$$

6.7 The behavior of the $\tilde{G}_k$ as shown in G for times $k \geq 50$ illustrates an important feature of the kind of system being considered. It will be noticed in Figure 6 that the number of operational craft present in $\Sigma^{\#}$ is considerably in excess of that called for by the Program IIa as given in (6.3.1).

The resulting excess is significant in that the system of inputs, (6.6.1), barely maintains the required operational strength until time 50. Essentially, the program (6.3.1) is too abrupt in its terminal stages. The system has a certain amount of "momentum" which cannot be overcome without "negative inputs" such as administrative abandonment or scrapping.

7. PROGRAM IIb

7.1 We shall now consider a more complicated kind of program starting with a non-empty $\Sigma^{\#}$ and converting to a different kind of plane during the program. A certain function is

being performed by a given kind of plane, and it is desired to keep the operational strength available for this function at the levels given in (7.1.1) below.

(7.1.1)

Strength Desired at Time $k = G_k$	k
50	0
60	1
70	2
80	3
90	4
100	$5 \leq k \leq 20$
105	21
110	22
115	23
120	24
125	25
130	26
135	27
140	28
145	29
150	$30 \leq k \leq 60$
145	61
140	62
135	63
130	64
.	.
.	.
.	.
0	90
0	$91 \leq k \leq 200$

For computational brevity it was decided to have planes in only two states of $\Sigma^{\#}$ at time 0 when we shall suppose that there are 30 aircraft in S_5 and 20 in S_{11} and no others anywhere in $\Sigma^{\#}$, save possibly in S_{31}. The treatment of the case where more states contained planes at the outset can be handled in much the same way and involves only additional computation.

The ships in S_5 and S_{11} at time 0 can be expected to contribute to the future operational strength of the system, so we must first compute these contributions and subtract them from the G_k of (7.1.1) in order to obtain the residual program to be supported by inputs into S_1. In order to accomplish this we first set

$$a_n = \sum_{j=1}^{8} p_{5j}^n + \sum_{j=12}^{19} p_{5j}^n + \sum_{j=23}^{30} p_{5j}^n$$

(7.1.2)

$$b_n = \sum_{j=1}^{8} p_{11,j}^n + \sum_{j=12}^{19} p_{11,j}^n + \sum_{j=23}^{30} p_{11,j}^n,$$

and then compute $K_n = 30a_n + 20b_n$. The quantities a_n and b_n will be recognized as the analogs of the ϕ_n for states 5 and 11, i.e., they are just the probabilities that a ship in $S_5(S_{11})$ at time 0 will be operational at time n. The quantities K_n are then the expected contribution to the operating strength of $\Sigma^{\#}$ resulting from the planes in S_5 and S_{11}.

The quantities K_n were computed and were found to be as follows:

(7.1.3)

n	K_n	n	K_n
0	30	14	19
1	46	15	13
2	39	16	12
3	33	17	11
4	15	18	14
5	18	19	11
6	19	20	7
7	35	21	6
8	31	22	5
9	23	23	3
10	21	24	2
11	21	25	2
12	25	$n \geq 26$	0
13	22		

Subtracting K_n from G_n we obtain the residual program to be supported. The quantities $\bar{G}_n = G_n - K_n$ are listed below.

(7.1.4)

n	$\bar{G}_n$	n	$\bar{G}_n$	n	$\bar{G}_n$	n	$\bar{G}_n$
0	20	15	87	$30 \leq n \leq 60$	150	75	75
1	15	16	88	61	145	76	70
2	31	17	89	62	140	77	65
3	47	18	86	63	135	78	60
4	75	19	89	64	130	79	55
5	82	20	93	65	125	80	50
6	81	21	99	66	120	81	45
7	65	22	105	67	115	82	40
8	69	23	112	68	110	83	35
9	77	24	118	69	105	84	30
10	79	25	123	70	100	85	25
11	79	26	130	71	95	86	20
12	75	27	135	72	90	87	15
13	78	28	140	73	85	88	10
14	81	29	145	74	80	89	5

with $\bar{G}_n = 0$, $90 \leq n \leq 200$.

Solving the equations $\sum_{j=0}^{k} \phi_j N_{k-j} = \bar{G}_k$ as before, the required inputs were found to be as follows:

(7.1.5)

k	N_k	k	N_k	k	N_k	k	N_k
0	20	21	13	42	12	63	6
1	0	22	15	43	12	64	6
2	17	23	16	44	12	65	5
3	20	24	16	45	13	66	5
4	33	25	17	46	13	67	4
5	17	26	18	47	13	68	5
6	8	27	18	48	13	69	3
7	0	28	18	49	14	70	3
8	4	29	21	50	14	71	2
9	11	30	20	51	14	72	4
10	10	31	15	52	15	73	2
11	4	32	14	53	14	74	3
12	11	33	13	54	14	75	3
13	7	34	12	55	15	76	3
14	4	35	12	56	14	77	1
15	2	36	12	57	14	78	2
16	5	37	12	58	14	79	1
17	9	38	12	59	13	80	1
18	7	39	11	60	13	81	1
19	10	40	10	61	8	$n \geq 82$	0
20	9	41	10	62	8		

The comparison between the proposed program $\{\bar{G}_k\}$ and that achieved under the system of inputs (7.1.5) is of particular interest here. The planes in S_5 and S_{11} just before time 0 serve to produce a surplus of operational craft twice in the early stages of the program. This happened because of the need for a negative N_1, i.e., the correct solution of $N_1\phi_0 + N_0\phi_1 = \bar{G}_1$ was less than zero. In the models being considered such N_j are taken to be zero, and a program in excess of $\{\bar{G}_k\}$ was induced. Also, the slope used during the latter stages of IIb, i.e., a drop of 5 planes each month, is shown to adhere much more closely to the natural rate of decay of operating strength than was the case in IIa where decrements of 10 were used. The comparison is shown in Figure 7.

7.2 Suppose now that a new aircraft, y, will be made available in certain quantities during the program IIb and that one desires to abandon the former model, call it x, as rapidly as possible. Then one proceeds to modify the N_k listed in (7.1.5) by reducing them as much as possible in the light of the availabilities for y.* For instance, suppose model y would be available in quantities R_k at time k where

*We are assuming here that the new model has the same p, q, r as x. More about this question below.

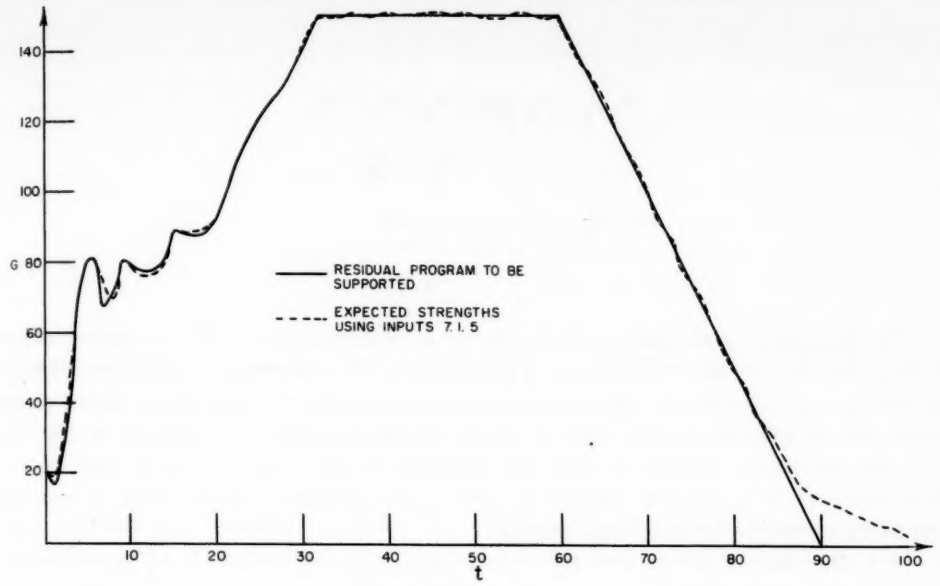


Figure 7

(7.2.1)

$$R_k = 0 \qquad 0 \le k \le 50$$
$$R_k = 1 \qquad 51 \le k \le 55$$
$$R_k = 2 \qquad 56 \le k \le 65$$
$$R_k = 5 \qquad 66 \le k .$$

Then if we let $\bar{N}_k$, $\bar{M}_k$ stand for number of model x to be used and for the maximum number of model y which can be introduced subject to IIb, we get the following schedule of inputs:

(7.2.2)

k	$\bar{N}_k$	$\bar{M}_k$	k	$\bar{N}_k$	$\bar{M}_k$	k	$\bar{N}_k$	$\bar{M}_k$	k	$\bar{N}_k$	$\bar{M}_k$	k	$\bar{N}_k$	$\bar{M}_k$	k	$\bar{N}_k$	$\bar{M}_k$
0	20	0	14	4	0	28	18	0	42	12	0	56	12	2	70	0	3
1	0	0	15	2	0	29	21	0	43	12	0	57	12	2	71	0	2
2	17	0	16	5	0	30	20	0	44	12	0	58	12	2	72	0	4
3	20	0	17	9	0	31	15	0	45	13	0	59	11	2	73	0	2
4	33	0	18	7	0	32	14	0	46	13	0	60	11	2	74	0	3
5	17	0	19	10	0	33	13	0	47	13	0	61	6	2	75	0	3
6	8	0	20	9	0	34	12	0	48	13	0	62	6	2	76	0	3
7	0	0	21	13	0	35	12	0	49	14	0	63	4	2	77	0	1
8	4	0	22	15	0	36	12	0	50	14	0	64	4	2	78	0	2
9	11	0	23	16	0	37	12	0	51	13	1	65	3	2	79	0	1
10	10	0	24	16	0	38	12	0	52	14	1	66	0	5	80	0	1
11	4	0	25	17	0	39	11	0	53	13	1	67	0	4	81	0	1
12	11	0	26	18	0	40	10	0	54	13	1	68	0	5	$n \ge 82$	0	0
13	7	0	27	18	0	41	10	0	55	14	1	69	0	3			

In other words,

$$\text{if } N_k \geq R_k \text{ then } \bar{N}_k = N_k - R_k$$
$$\bar{M}_k = R_k ,$$
$$\text{if } N_k < R_k \text{ then } \bar{N}_k = 0$$
$$\bar{M}_k = N_k .$$

7.3 It is fairly clear that the above method of replacing x's by y's is meaningless when the system parameters for the two kinds of plane differ. By system parameters are meant the quantities p, q, r and whatever other probabilities may be basic to a description of the system, together with the various numbers describing such things as length and frequency of operational periods and overhauls. In order to show how the situation may be handled in the case of two kinds of plane x and y with two systems Σ_x and Σ_y determining their behavior, we consider the following variant on the preceding example.

7.4 The plane y will be available in quantities y_j, $j \geq 0$, and let ϕ_k^x, ϕ_k^y correspond (for the two aircraft types) to the numbers ϕ_k as defined above. One proceeds in a sequence of steps σ_j as follows:

σ_0: Solve $N_0^y \phi_0^y = G_0$ for N_0^y

$$\text{Set } \bar{N}_0^y = \begin{cases} 0 & \text{if } N_0^y < 0 \\ N_0^y & \text{if } 0 \leq N_0^y \leq y_0 \\ y_0 & \text{if } N_0^y > y_0 , \end{cases}$$

and set $\bar{y}_1 = y_1 + y_0 - \bar{N}_0^y$.

σ_1: Solve $\bar{N}_0^y \phi_1^y + N_1^y \phi_0 = G_1$ for N_1^y.

$$\text{Set } \bar{N}_1^y = \begin{cases} 0 & \text{if } N_1^y < 0 \\ N_1^y & \text{if } 0 \leq N_1^y \leq \bar{y}_1 \\ y_1 & \text{if } N_1^y > \bar{y}_1 \end{cases}$$

and set $\bar{y}_2 = y_2 + \bar{y}_1 - \bar{N}_1^y$.

σ_2: Solve $\bar{N}_0^y \phi_2^y + \bar{N}_1^y \phi_1^y + N_2^y \phi_0^y = G_2$ for N_2^y

$$\text{Set } \bar{N}_2^y = \begin{cases} 0 & \text{if } N_2^y < 0 \\ N_2^y & \text{if } 0 \le N_2^y \le \bar{y}_2 \\ \bar{y}_2 & \text{if } N_2^y > \bar{y}_2 \end{cases}$$

and set $\bar{y}_3 = y_3 + \bar{y}_2 - \bar{N}_2^y$.

Proceeding in this way we obtain a system of inputs $\bar{N}_k^y$ for aircraft type-y. Next the quantities

$$\tilde{G}_k = \sum_{j=0}^{k} \bar{N}_{k-j}^y \phi_j^y$$

are computed and the "residual" program $\{G_k - \tilde{G}_k\}$ determined. Finally the equations

$$\sum_{j=0}^{k} N_j^x \phi_{k-j}^x = G_k - \tilde{G}_k$$

are solved in the usual way for the inputs of the x-type plane.

7.5 The introduction of the quantities $\bar{y}_k$ in the above serves simply to refurnish at time k+1 any surplus of y existing just after time k.

8. REMARKS ON THE N_i

8.1 The behavior of the inputs N_i is worthy of special note in at least one case. Suppose that a system like $\Sigma^{\#}$ is empty prior to time 0 and that the program of operating strengths to be supported has the form $G_j = \lambda$, $0 \le j$, λ a constant. The associated N_i can oscillate rapidly, aperiodically, and for a considerable period of time. In a trial computation the inputs were observed to rebound in an irregular manner from limits roughly on the order of $\pm 0.8\lambda$ during the early stages of the program and to level off only after a period of years. Thus, overly rapid build-up programs can be expected to have rather striking effects upon the numbers of planes in operational status.

9. FURTHER THEORETICAL REMARKS

9.1 Suppose that one has a program $\{G_k\}$ to be met in a system described by a matrix P like the matrix $P^{\#}$ used above. We shall imagine that the program is to be continued indefinitely and that from some time on, say k+1, the G_i are all the same, i.e., $G_{k+j} = \lambda$, $j = 1,2,3,\ldots$. One might inquire as to what extent the solution to the system

$$\sum_{j=0}^{k} N_{k-j} \phi_j = G_k \quad k = 0, 1, 2, \tag{9.1.1}$$

may be expected to resemble the G_k. Specifically, will the N_j be constant or almost constant for all sufficiently large j? In order to answer this question we define formally the function N(x) as follows:

(9.1.2) $$N(x) = \sum_{k=0}^{\infty} G_k x^k \Big/ \phi(x)$$

where $$\phi(x) = \sum_{k=0}^{\infty} \phi_k x^k .$$

In the cases at hand, say of $P^{\#}$ or, for that matter, any of those matrices P associated with actual aircraft types, $\phi(x)$ is a polynomial. ($\phi_k \neq 0$ for indefinitely large k would simply mean that no retirement provision has been made for the aircraft in question.) Using the above definition of the G_k we can therefore rewrite (9.1.2) as,

(9.1.3) $$N(x) = \frac{\sum_{i=0}^{k} G_i x^i + \lambda \sum_{j=1}^{\infty} x^{k+j}}{\sum_{j=0}^{r} \phi_j x^j} .$$

From (9.1.3) N admits expansion in a power series about the origin converging in the disk $o \leq r \leq a$ where $ae^{i\theta}$ is the root of $\phi(x)$ which is nearest the origin. Moreover, the multiplication

$$N(x) \cdot \sum_{j=o}^{r} \phi_j x^j = \sum_{i=o}^{\infty} N_i x^i \cdot \sum_{j=o}^{r} \phi_j x^j$$

$$= \sum_{j=o}^{\infty} \left(\sum_{i=o}^{j} N_{j-i} \phi_i \right) x^j = \sum_{j=o}^{\infty} G_j x^j$$

must be valid when $|x| < a$.

Accordingly,

(9.1.4) $$\sum_{i=o}^{j} N_{j-i} \phi_i = G_j .$$

Since the solution to (9.1.4) is unique, the coefficients N_j of the power series for N(x) must coincide with the inputs required to support the program $\{G_j\}$.

From (9.1.3) we have

$$\sum_{i=o}^{\infty} N_i x^i = \frac{\sum_{i=o}^{k} G_i x^i}{\phi(x)} + \frac{\lambda x^{k+1}}{(1-x)\,\phi(x)}$$

or

$$(1-x) \sum_{i=o}^{\infty} N_i x^i = \sum_{i=o}^{\infty} (N_i - N_{i-1}) x^i$$

$$= \left[\sum_{i=o}^{k+1} (G_i - G_{i-1}) x^i \right] \Big/ \phi(x)$$

where $N_{-1} = G_{-1} = 0$. By definition of a we have

$$\limsup_{i\to\infty} \left|N_i - N_{i-1}\right|^{\frac{1}{i}} = \frac{1}{a}.$$

In other words, if $\phi(x)$ has a root in the unit disk, the N_j can be expected to oscillate with ever increasing amplitude. If $a>1$, the N_j converge to some limiting N. In the case $a = 1$, the N_i will eventually oscillate rather narrowly with an amplitude of about 1.

Sections 7 and 8 call attention to the questions of utility to which one is led in the kind of situation under study. The technique proposed in Section 7 for handling more than one kind of plane settles little, save under very restrictive assumptions upon the availabilities of the two types of aircraft. Similarly, the questions which come to mind in connection with paragraph 8.1 can be approached only in the presence of a lot of quantitative information and settled only when decisions have been made concerning the military utility of various procedures. The system problems, large groups of inoperative aircraft, sudden heavy demands upon repair facilities, etc., must be weighed against the need for rapid increases of operating strength.

These questions are all important and should be studied. It is hoped that the present work may contribute to that purpose through its description of some of the underlying phenomena.

BIBLIOGRAPHY

[1] Feller, W. An Introduction to Probability Theory and Its Applications, Volume I, 1950, John Wiley and Sons, Inc.

[2] Loève, M. Probability Theory, 1955, D. Van Nostrand Company, Inc.

[3] Kelley, J. "Calculation of the Aircraft Support Problem on the Logistics Computer," 1955, Serial 3374, Logistics Research Project of The George Washington University.

* * *

THE CONCEPT OF A "MISSION"

Michel Verhulst[1]
Princeton University

The author discusses the concept of a mission within the framework of the Theory of Games and analyzes a simple model.

Few words are more often used in military circles than the word "mission." But few words are more difficult to define precisely. One speaks of the mission of the U. S. Air Force, of the mission of a Naval Supply Depot, of the "operation of an actual support mission," of the mission of a logistics conference, etc. The word is so much taken for granted that it might even appear naive to devote a full paper to the concept of a "mission." In fact, the word covers so many different "objective functions," to use the jargon of the analyst, that the present paper studies only one very simple type of mission. But what it loses in breadth of vision, it wins in depth of understanding.

PART I

DEFINITION OF A MISSION

The official military definition of a "mission" as given in the "Dictionary of United States Military Terms for Joint Usage" is the following:

> Mission. - 1. The objective; that is, the task together with its purpose, thereby clearly indicating the action to be taken and the reason therefor. 2. In common usage, especially when applied to lower military units, a duty assigned to an individual or unit; a task. 3. The dispatching of one or more aircraft to accomplish one particular task.

It is clear that these different meanings have some basic elements in common and in the present paper we will say that a mission implies the selection by an authority of an objective, usually well defined and which must be reached within a certain time period, through a chain of operations to be chosen and performed by the agents in charge of the mission and responsible for its successful conclusion, in spite of eventual difficulties and obstacles which cannot be well defined at the beginning of the operations.

To take a simple example, the obligation put upon me to pick up a car at a factory before closing time on Friday, knowing that I can leave my work only on Thursday night and that the factory is 650 miles away, is a mission. In this case, the time period is precisely defined and the mission has none of the characteristics of a flow through time. This is true also in the case of any type of flight of the Military Air Transport Service of the U. S. Air Force. One type of flight, for instance, might be the obligation to carry 60 passengers from MacGuire Air Force Base to Frankfurt-am-Main. The flight should depart at a certain time, arrive at a specific time, and stop in between, if necessary, for refueling.

---

[1]This paper has been written at Princeton University under Office of Naval Research Contract N6onr-27009.

But in the more sophisticated case of logistics planning, the "flow" character is essential. For a Naval Supply Depot, for instance, a mission can be defined as the obligation to meet or fulfill a general demand for certain quantities of products. Under each mission, however, will normally come a series of requisitions or orders for products, ranging from daily orders to replenishment orders which may be many months apart. In all such cases, the necessity of associating a time reference to the mission and the requisitions is obvious. But the best way of doing it is a matter for discussion.[2]

In the present study, then, the concept of order will be considered as essentially the same as the concept of mission defined above, so that the concept of mission as used by the Navy covers really a flow of missions, in the narrow sense in which we use this word. It seems to us, however, that the restricted definition is useful for the purpose of the present paper because it is very precise and puts the emphasis on the functional aspect of the logistical situations. This aspect is too often underestimated in the current literature on logistics, although it provides the key to the solution of many apparently complicated problems. This does not mean, however, that the macroscopic approach to the problems of logistics is not useful. In fact, we should tackle eventually the aggregation problem of reconciling our microscopic approach with the macroscopic one, since it is only at this latter level that suggestions or recommendations could be of real use.

BASIC CHARACTERISTICS OF A MISSION

In the definition of a mission, particular emphasis must be placed on the notion of selection of the objective. This implies that the objective selected at a given time by the Authority has priority on other objectives and, therefore, has also priority on the use of the resources at the disposal of the Authority. But the difficulties and the obstacles that the actual realization of the mission may encounter, and also the unavoidable changes in priorities which come with the evolution of dynamic situations, have an influence on the outcome of the mission and on its real value when it is completed. A distinction also must be drawn between the technical feasibility of a mission at any cost and its feasibility within a maximum cost. The degree of priority given to the mission has a definite bearing on this problem. Moreover, the cost in question might not only be a money cost, but also a kind of psychic cost where the Authority and the agents have to force themselves, say, to analyze the relevant information which is available, perhaps at no cost in money, but supposing, on their part, in order to be useful, the will to think and plan. Also, once the mission has started, the penalty of failure[3] might be such that the Authority and the agents may be willing to forget altogether about cost when it is realized, in the course of the attempt to fulfill the mission, that the probability of failure has increased considerably for one reason or another not foreseen at the beginning of the operations. The interrelationships between all these factors should be studied in more detail, and we shall try to do so in the course of the present paper.

Another essential character of a mission is the notion of operations and of the mathematical sequence or order in which they must be performed. A mission supposes a set of operations which must be performed usually in a certain order[4] but which have alternatives

[2] See O. Morgenstern, "Consistency Problems in the Military Supply System," Naval Research Logistics Quarterly, Vol. 1, No. 4, December 1954.

[3] This concept will be defined more precisely in the course of the paper.

[4] There are, however, certain subsets of operations which can be performed indifferently in one way or another: filling a container, for instance, with products coming under the same order in a Naval Supply Depot.

involving different costs and which correspond to different courses of action, each course being any set of successive operations which leads to the successful completion of the mission. The completion of each successive operation opens a new subset of operations, one of which is selected in such a way that it entails, say, a minimum cost (money cost and psychic cost), given the information available and the risk of ultimate failure that can be accepted at any given time within the time period which has been fixed.

Naturally, in order to bring the mission to its successful conclusion, some organization must be set up which must be adapted to the needs of the particular mission. This organization will draw on the environment and will make use in a rational way of the resources available, given the cost and the probable penalty constraints which are relevant at the particular point of time reached within the time period.

IMPORTANCE OF THE THEORY OF GAMES FOR THE STUDY OF MISSIONS

In many logistical situations, a mission can be assimilated to a move in the sense defined in the theory of games originally developed by O. Morgenstern and John von Neumann. In fact, the body of the Authority and of the corresponding Agents can be assimilated to a player in a kind of strategic game which could be called Race against Time, and in which a particular mission, when decided on, represents a move in the game. Naturally, in time of war, the other players of the game would be the bodies of the opponent authorities and of their agents who try directly or indirectly to nullify the expected results of the mission and who have missions of their own. Their corresponding moves involve also the use of Time as a kind of battlefield, the possession of which, or, rather, the most effective use of which, is vital for the final issue and the ultimate value of the operations. But the use of Time involves the use of other resources at a rate which can be measured relatively accurately and even forecasted in many cases by means of, say, usage data. The problem of the best use of Time is, therefore, closely related to the problem of the optimum use of resources. But for our present purpose, and since Time appears as the most valuable commodity to which all others are related, we can refer to the game visualized here as the game of the Race against Time.

In time of peace, the actions of the different bodies of the authorities and their agents, i.e., the moves of the players, may not be quite as obvious as in time of war. But whether the logistical situations are brought about by the necessity either of keeping acquired positions or of attempting to win new ones, the essential structure of the game is still the same. Admittedly, the importance and the number of moves decided on per time unit are vastly different in time of war or in time of peace, so that the dynamics of the game may be quite different according to the volume and the scale on which the operations are conducted. In fact, the problem of expansion in war should deserve more attention than it receives generally, along with the problem of rapid identification and special handling of critical items in war.[5]

Such a general view of the meaning and purpose of logistical situations considered as a whole, however, is very often insufficient to provide good rules of action in particular situations. It is only at a certain level of abstraction of the over-all process of decision-making that methods of analysis and of behavior based on this approach could be helpful. This explains why, in particular logistical situations where one move of the game is all that counts for the agents put in charge of a particular mission, other methods of rational behavior for the agents and of control for the authorities must be found which must be sound from the general point of view of the theory of games but also practical from the point of view of everyday use.

---

[5] Remark of Rear Admiral H. E. Eccles to the author.

Still, at this lower level of analysis, it seems reasonable to assume that to win means to fulfill the mission in the time allocated. But Time cannot be considered as a conscious human being who tries to thwart systematically the efforts of the Agent. It is just a name to cover what appears to be, for all intents and purposes, the substratum of the moves of the players in the game of the Race against Time as well as the stake of the game. A scale of values, however, should be devised in most cases to weigh the importance of Time according to the type of missions and of logistical situations involved.

Now, if the mission could be fulfilled as planned, with complete certainty, and if the information were complete and perfect, there would be no real problem, because the decision to make a move and the actual implementation of this move would be identical. The theory of games would then be fully adequate to provide methods of rational behavior in all cases. In fact, we would know in advance that the mission is technically and economically feasible, and all operations which constitute the mission could be determined before the start. There might even be different courses of action involving different costs, but they would all lead to the successful conclusion of the mission, and the cheapest course would be the best, provided the time constraints were satisfied. In short, this would be planning, forecasting, and action under ideal circumstances.

In practice, however, to enlarge the horizon of planning costs money and effort. Very often all that can be done at first by the agents is to obtain a broad idea of the kind of operations involved and of their rough timing, and to rely on the existence of possible alternatives in case something goes wrong with the course of action which appears to be the cheapest. This means that the horizon of definite planning or, should we say, of relatively accurate forecasting, is shorter than the duration of the mission. In fact, it extends very often only to the first operation of the mission, because of the possibilities of direct or indirect interference of the other players in the game, and also through the impossibility of obtaining all relevant information. Then, if the first operation is completed within a certain time—if I get to the intermediate station in time to catch the next train, for instance—all is well and we can say that Time has lost the first round of the actual implementation of the move decided on by the Authority. If it is not the case, Time has won, and the penalty is that the Agent loses the use of the cheapest course of action and must now embark on another course of action, more expensive than the first one. Then the effort to make the move starts again with another round, the Agent taking the course of action which is the least costly among those which in his opinion will bring about the ultimate success of the mission. And so on, until the end. The final success of the mission, i.e., the actual implementation of the move decided on, is achieved when it is known whether Time has been defeated for sure.[6]

THE CONCEPT OF THE VALUE OF A MISSION

One fact appears immediately to be important in the definition of the value of a mission. It is to know whether the mission once decided has been started and whether it is now too late to back out and to adopt another mission which might be better if it had to be decided all over again. In the latter case it means, presumably, that the cost of the mission already incurred up to the present time is too sizable, so that to start a new mission at this point would be too costly with respect to the expected supplementary advantage that could be obtained by embarking now on another mission. In fact, it can be assumed in many cases that to order the Agent to switch to

[6] In the case of missions such as those given a Naval Supply Depot, it would also be necessary that the right quantities of products required had been shipped and received at their place of use.

another kind of mission when one is already developing would not bring a supplementary advantage, but that the equivalent of the planned advantage visualized at the start of the mission can always be obtained by changing the planned course of action of the present mission: adopting, say, a more expensive course of action when it is realized that the expected breaks do not materialize or that the enemy has started a counteracting action of his own in answer to our mission. In this sense, the present mission may be turned gradually into another kind of mission, but we can discard this fact and say, for all intents and purposes, that there is only one optimum mission attached to a given horizon of planning and that the present mission, once completed, is that one. This optimum mission, however, is defined at the start of the mission in terms of the a priori probabilities of the different courses of action and of the corresponding operations which compose the mission, these probabilities being really the subjective probabilities attached by the Authority to the probable actual realization of the operations. These probabilities, although subjective, have nevertheless a great importance, since it is on them that the Authority is prepared to make its decision as to whether the present mission is the best or not, and to bet its reputation of being fit for the job of taking this kind of decision. They can be looked at as the outcome of innumerable experiences similar to the one under development. Theoretically, the possible choices between the alternative operations at each stage of the mission are potentially very great, since they are determined partly by the actions of the enemy who may have started a mission of his own to counteract the effects of our own mission and partly through chance, since there are some breaks that we can reasonably expect but which may not materialize. In practice, however, there are but a few possible operations and courses of action which really matter, and those are the only ones retained by the Authority in making its decision. These operations and courses of action are those which would appear again and again, and have the higher frequency, if the same kind of logistical situation and the same kind of mission would be found in other places and other times. This set of operations, together with their timing and the probabilities attached by the Authority to their realization in time, is the operational definition of the concept of a mission.

Under certain assumptions, we can then give a more concrete meaning to the concept of the value of the mission. The expected pay-off or expected value of the mission is the difference between the expected value of the advantage of the mission and its expected cost. Let us show first how to compute the latter.

THE EXPECTED COST OF A MISSION

The chances of getting some breaks in the actual development of the operations of the mission can be taken care of very simply by imagining that the expected cost of an operation, once decided, can take different values, making this move more expensive than average, less expensive than average, or normally expensive, with given probabilities of realization in all three cases. Let us assume that these probabilities are all equal from the standpoint of the Authority.

The different operations of the mission, however, do not have the same probability of realization from the standpoint of the Authority. Some courses of action have a greater probability of realization than others and, naturally, the probabilities of realization attached to each last operation in the different potential courses of action are identical to the probabilities of these courses of action. If we assume that these latter probabilities are known, the probabilities of all the operations in each potential course of action also are known. It is then easy to compute the expected cost of the mission by weighing the average cost of each operation by the

corresponding probability of the course of action to which it belongs and adding up the weighted costs of all the potential operations. Naturally, we could also compute the average cost of the potential courses of action and weigh them by their respective probabilities of realization. The sum of these costs would also be the expected cost of the mission.

THE EXPECTED ADVANTAGE OF A MISSION

In many cases, instead of using the direct method based on the assimilation of the broad logistical situation to a game and of the mission to a move in the game, it seems possible to use an indirect method for the purpose of computing the value of the advantage of the mission. In fact, if we make the assumption that the reason of a move is to force the enemy to make a corresponding move and to pay for his own move more than it costs us to make our own move, we have then a simple way to measure the advantage of a move: this advantage is equal to the cost of the move of the opponent.

Naturally, the expected cost of the move of the opponent is not so well defined, usually, as the expected cost of the mission under consideration by the Authority. Experienced commanders, however, can guess this cost well enough in routine situations, especially when the mission is sufficiently broad in scope. They may even guess this cost better than the opposite commanders themselves when their information on usage data and their knowledge of the ways to use it for planning purposes is better than those at the disposal of the enemy. But this method assumes implicitly that the Authority has the initiative of the move and this question deserves an explanation.

DEFINITION OF THE CRITICAL MOVE

In broad strategic situations, there are always many moves which take place more or less at the same time, and it is difficult to single out one particular move as being definitely responsible for a certain move of the opponent. To a certain extent, the preceding analysis has gone around this difficulty by assuming that a mission can be turned gradually in time into another one, and this is really a way to take care of a chain of moves such as those met in real strategic situations. This allows us to match the chain of moves of the Authority against the corresponding chain of moves of the opponent. As long, then, as the Authority is able to make the opponent pay more for his moves than what it costs for its own moves, we say that the Authority keeps the initiative. If this trend if reversed, the Authority loses the initiative. Now, to be in the position of having the initiative is important since it implies that the general development and the scale of the operations are at the disposal of whoever has the initiative. The move, whether it is a mission or an operation, which corresponds to that position where it is necessary to let the initiative go to the opponent can be called the critical move. Naturally, the critical cost is the actual cost of this critical move. It is also the actual cost of the move of the opponent under the assumption made above that the advantage of a move is equal to the cost of the move of the opponent. It means also that the critical cost is equal to the respective advantages of both moves and that the expected pay-off of the mission is zero or close to zero when the mission is critical.

DEFINITION OF THE PENALTY OF FAILURE

This allows us to give a precise meaning to the concept of the penalty of failure. We can say that it is the discounted value of all the losses that the Authority will suffer at the hands of the opponent because of its failure to keep the initiative at a critical point in time, a loss being

the difference between the cost of a mission and the corresponding advantage when the Authority does not have the initiative.

Naturally, this concept is general and can be applied to all sorts of strategic situations. In this paper, we restrict it to the set of a mission and of the successive missions that will follow the first one, perhaps against the will of the Authority because the mission under consideration has become critical. Eventually, the Authority may regain the initiative either directly, or, more probably, indirectly by attacking the enemy on another field and forcing him to lose the initiative. But this is another problem which cannot be considered in the present paper. It also might happen that the chain of moves which will follow the critical move will come to a natural end. The concept of the penalty of failure can therefore be very useful since the penalty can be computed most of the time, not always on a priori grounds except as an estimate but certainly a posteriori, when the chain of moves has come to an end.

In the simple example given above of the mission which consists of picking up a car at the factory before closing time on Friday, it is the last operation of the mission which may become critical, i.e., the necessity, say, of catching a plane instead of taking the train as planned at the beginning, since we had missed the train connection at the intermediate station. If the fare of the plane had been such that it would not have been worth while getting to the factory on time but, rather, to proceed leisurely by train to the factory and to suffer the consequences, it would have meant that the extra cost of taking the plane would have been equal to or greater than the penalty of failure. The penalty of failure appears, therefore, as the price of failing to keep the initiative in time and of not meeting the time requirements of the mission as given at the start of the mission.

PART II

A SIMPLE MODEL OF A MISSION[7]

To illustrate the results obtained so far, we can draw a graph showing the evolution of the cost of the typical mission described above as a function of time, assuming that the value of the advantage of the mission has not changed during the course of the mission (Figure 1).

The Authority has to decide whether the mission is worth undertaking and the two courses of action which seem possible are as follows: either the Agent will go by train all the way or he will go by train all the way up to the last leg of the journey, for instance, and will then take a plane to complete the mission. But it is not certain whether a plane will be available.

For the first alternative, the cumulative cost curve of the operations

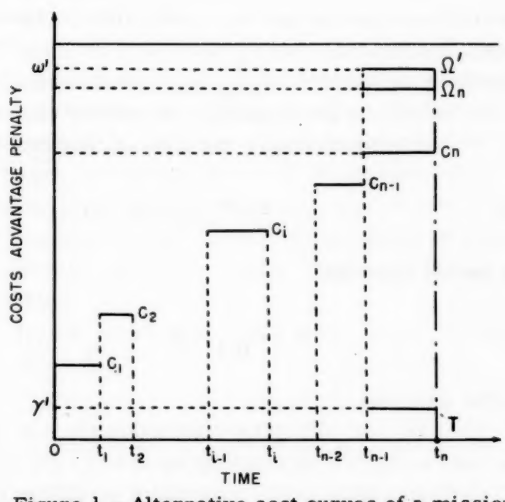


Figure 1 - Alternative cost curves of a mission compared to the advantage of the mission

0T = time allocated for the mission
0ω' = value of the advantage of the mission
ω'γ' = penalty of failure if the mission is not completed on time
C_1, C_2, C_i, C_n = normal cost curve of the mission
C_1, C_2, C_i, Ω_n = alternative cost curve of the mission

[7] I am indebted to A. J. Hoffman for his advice on this point, and also to Michio Hatanaka.

performed at times $t_1, t_2, \ldots, t_n$ ends up at the point C_n on Figure 1. For the second alternative, it ends up at the point Ω_n, and the segment $C_n\Omega_n$ represents the extra cost of taking the plane instead of going by train on the last leg of the journey. It is the last operation of the mission which may become critical in this case.

The value of the advantage of the mission is represented by the ordinate $0\omega' = T\Omega'$. It is the difference between the price of the car on the local market and the price charged at the factory.

FIRST PROBLEM

Assuming that the intermediate point where the Agent must make a decision concerning whether to switch from the train to the plane is known (the beginning of the last leg of the journey, for instance), and given the subjective probability π in the mind of the Authority that the Agent will try to catch a plane when he reaches this intermediate point because he will not be certain otherwise to complete the mission on time, what is the condition which determines whether or not the Mission is worth undertaking? We assume, also, that the Agent will actually complete the mission on time if he decides not to try to catch a plane because he feels certain to complete the mission on time.

Let ρ be the probability in the mind of the Authority that the Agent will actually miss the plane if he tries to catch it. Naturally, if the Agent does not catch the plane, the mission cannot be completed on time but the Agent must finish it all the same by train.

There are two cases to consider according to whether the Authority is certain (1st case) or not certain (2nd case) that the mission will be completed on time if the Agent does not miss the plane. In the 2nd case, let ε be the probability that the mission will not be completed on time. Presumably, ε is very small.

Let A be the value of the advantage of the mission and C_{n-1} the cumulative cost of the operations up to the last leg of the journey. Let also γ_n be the cost of the last operation if the Agent goes by train and γ_a the cost if the Agent takes a plane; the difference $\gamma_a - \gamma_n$ is represented by the segment $C_n\Omega_n$ on Figure 1. Finally, let F be the value of the penalty of failure if the mission is not completed on time for one reason or another.

Before undertaking the mission, the expected cost of the mission is:

$$\Delta = C_{n-1} + (1 - \pi)\gamma_n + \pi\,[(1 - \rho)\,\gamma_a + \rho\,(\gamma_n + F)] \tag{1}$$

in the 1st case, and

$$\Delta = C_{n-1} + (1 - \pi)\,\gamma_n + \pi\,[(1 - \rho)\,(\gamma_a + \varepsilon F) + \rho\,(\gamma_n + F)] \tag{1}'$$

in the 2nd case.

The term in brackets expresses the fact that, if the Agent does not miss the plane when he tries to catch it (probability equal to $1 - \rho$), the expected cost of the last operation is γ_a or $\gamma_a + \varepsilon F$, according to the case, and if he misses the plane (probability ρ) and goes instead by train although he knows that he should have taken the plane, then he pays the cost γ_n and the penalty F.

Reasonable values for π and ρ might be 3/4 and 1/4, for instance, and the value of ε might be 0.05.

Now, the mission is worth undertaking only if the value of the advantage of the mission is greater than the expected total cost of the mission, i.e., if:

$$A - \Delta \geq 0 \,. \tag{2}$$

This condition says also that the expected pay-off of the mission must be greater than or equal to zero. It is the condition we are looking for.

SECOND PROBLEM

At which intermediate point should the Agent try to switch to a plane? It is important that the Authority should tell the Agent at which point he has to switch to a plane if he is not sure to complete the mission on time when he reaches that point.

There are several factors which influence the determination of the optimum intermediate point, and a proper balance has to be found between their values. This balance is found when the expected pay-off of the mission is as great as possible or, what is the same, when the expected cost of the mission is as small as possible. Let us then consider the different parameters which have an influence on the value of the expected cost defined, for instance, by equation (1) written as follows:

$$\Delta = C_{n-1} + \gamma_n + \pi\,(\gamma_a - \gamma_n)\,(1 - \rho) + \pi\rho F \tag{3}$$

i.e.,

$$\Delta = C_n + \pi\,(\gamma_a - \gamma_n)\,(1 - \rho) + \pi\rho F \tag{4}$$

where C_n is the cost of the trip if it is made entirely by train.

In this equation, we see that the probability $\pi(t)$ varies according to the intermediate point[8] considered, because it depends on the estimate that the Authority makes of the chance events which may happen between the beginning of the mission and the time when the Agent reaches the intermediate point. This probability depends, for instance, on the chance of making good train connections at intermediate points before the point considered. In the same way, the probability $\rho(t)$ varies also as a function of time. It depends, for instance, on the eventual state of the weather, since the plane will not fly if the weather is bad, or on the availability of regular flights at some intermediate points. Finally, the value of $\gamma_a - \gamma_n$ varies in time. In fact, it could be approximated by a function g(t) defined by:

$$g(t) = -k\left(\frac{t}{T} - 1\right) \,. \tag{5}$$

where k is the value of the extra cost if the whole trip is made by plane and not by train and T is the total time allocated for the mission.

If we could assume that these functions are continuous and well defined over the time interval, we see that the expected cost would be minimum from the point of view of the Authority when the derivative of Δ as a function of time is zero, i.e., when

---

[8] We assume for the time being that the time scale and the point scale are the same. This is the reason of the symbols $\pi(t)$ and $\rho(t)$, where t is the time.

$$(6) \qquad [(1-\rho)\,g(t)+\rho F]\frac{d\pi}{dt}+\pi[F-g(t)]\frac{d\rho}{dt}+\pi(1-\rho)\frac{dg(t)}{dt}=0\,.$$

Taking into account the value of $\frac{dg(t)}{dt}=-\frac{k}{T}$ obtained from equation (5), this means that

$$(7) \qquad [(1-\rho)\,g(t)+\rho F]\frac{d\pi}{dt}+\pi[F-g(t)]\frac{d\rho}{dt}-\pi(1-\rho)\frac{k}{T}=0\,.$$

This condition determines the point of time where the Agent should try to switch to a plane. In practice, however, the functions $\pi(t)$ and $\rho(t)$ are not continuous and may not even be defined in a simple way at the beginning of the mission for the whole time interval. The condition which determines the equilibrium point is then more complicated. In fact, what we must determine is a good strategy, i.e., a method which tells us what to do in each case which may happen. This leads us to the third problem which defines a mission as a game where the player (the Agent) plays against chance. But we assume that the Agent knows the mixed strategy of chance, i.e., the probability distribution of the different strategies of chance. This means also that the probability distribution of the chance moves is known each time that the Agent makes himself a move. This generalization of the problem therefore takes into consideration the information obtained from the outcomes of the chance events so that the criterion for determining whether to change for a plane at time t, for instance, depends on the known outcomes of the chance events which took place before t (like the making of train connections).

THIRD PROBLEM

The mission considered as a one-person positional game.[9]

A mission in the sense considered in the present paper is a positional game where the player (the Agent) plays alternatively and successively with chance. The position x of the game is the topological product of the intermediate point s where the Agent finds himself, the corresponding time t, the amount of money m still available for the mission, and the means of transportation ε that he uses. The amount of money m is the difference between the total amount of money allocated by the Authority at the beginning of the mission and the amount which has already been spent by the Agent.

Let X be the set of all the possible positions of the game, and $\Gamma_1(x)$, $\Gamma_0(x)$ two multivalued functions from X to X; if, in a position x, the Agent has the move, $\Gamma_1(x)$ denotes the set of all the positions he can obtain theoretically at the next stage of the game; if, in the position x, chance has the move, $\Gamma_0(x)$ is the set of the possible positions which can occur effectively, with a known probability distribution $\{p(y)/y \in \Gamma_0(x)\}$.

The mission succeeds if the Agent is at a place s_0 before a time t_0, i.e., if the move belongs to the Agent, and if the position x lies in the set $K=\{(s, t, m, \varepsilon)/s=s_0;\ t\leq t_0;\ m\geq 0\}$. In this case, the game comes to an end, and the payoff is $f(x)=m$.

The mission fails if the move belongs to the Agent, and if the position x lies in the set $K'=\{(s, t, m, \varepsilon)/s=s_0;\ t>t_0 \text{ or } m<0\}$. In this case, the game stops, and the payoff is $f(x)=m-\mu$ where μ is the penalty of failure. The aim of the Agent is to obtain as high a payoff as possible.

---

[9] See J. Milnor, Contributions to the Theory of Games, Vol. II.

If $x_0 = (s_0, t_0, m_0, \varepsilon_0)$ is the position of the game, it means that the Agent is at the station s_0 at the time t_0 with an amount of money m_0, and uses the means of transportation ε_0 (a train for instance). If the Agent has the move, he can choose among several means of transportation $\varepsilon_1, \varepsilon_1', \varepsilon_1''$, etc.... A position of $\Gamma_1(x_0)$ will be $x_1 = (s_1, t_1, m_1, \varepsilon_1)$, where s_1 is the next station after s_0 for the train ε_1, t_1 is the expected time at which the train is supposed to arrive at this next station, and $m_0 - m_1$ is the fare from s_0 to s_1. If the next position is $x_2 = (s_2, t_2, m_2, \varepsilon_2) \in \Gamma_0(x_1)$, we shall have $s_2 = s_1$, $m_2 = m_1$, $\varepsilon_2 = \varepsilon_1$, and t_2 is the time at which the means of transportation ε_1 (the train, for instance) has effectively arrived at the station s_1.

A player i's strategy σ_i is a univalued function from X to X such that $\sigma_i(x) \in \Gamma_i(x)$ for all x. Let $z(\sigma_0, \sigma_1)$ be the position of $K \cup K'$ ("terminal position") obtained if the Agent plays with the strategy σ_1 and if chance "plays" with the strategy σ_0. If $p(\sigma_0)$ is the probability that chance will choose the strategy σ_0, then the Authority must choose a strategy σ_1 for which the expected payoff defined by

$$F(\sigma_1) = \sum_{\sigma_0} p(\sigma_0)\, f\,[z(\sigma_0, \sigma_1)]$$

is as high as possible. If the space X is finite, it is well known that there exists a good strategy.[10] If X is infinite, i.e., in the case where $\Gamma_0(x)$ or $\Gamma_1(x)$ are infinite sets of choices, there is not always a good strategy. We are going to show, however, that in the case where the mission can be represented by a topological game, there exists a definite good strategy and we shall give also a method to determine this strategy. These results have been found by C. Berge after a discussion with the author of the present paper.

FOURTH PROBLEM

Case where the mission is a topological game.[11]

Consider on a metric space S called "the sea" a moving point s called "the boat" which is trying to reach a point s_0. There are areas on the sea which are forbidden, such as islands, etc..., and areas which are dangerous because the boat might be sunk by the enemy or be incapacitated. The probability p(s, t) of this possible event ε is supposedly known at each place for the time t. We note that if the captain of the boat (the Agent) must take a decision every minute, for instance, then we have a mission in the sense defined above.

Let $x = (s, t, m, \varepsilon)$ be the position of the game, where s describes the place where the boat is, t is the time, m is the quantity of resources (fuel, etc...) still available, and ε is a variable which describes the degree of incapacity of the boat ($\varepsilon = 1$ if the boat is in perfect order, and $\varepsilon = 0$ if the boat is out of order). The set $\Gamma_1(x)$ is the set of all the positions of the game that the boat can reach in one minute from the position $x = (s, t, m, \varepsilon)$ supposing that ε has not changed during that time interval; if $\varepsilon = 0$, for instance, then $\Gamma_1(s, t, m, 0) = \{s, (t + 1), m, 0\}$. The set $\Gamma_2(x)$ is the set of the positions $x' = (s', t', m', \varepsilon')$ which can be reached effectively instead of the position $x = (s, t, m, \varepsilon)$, i.e., such that $s' = s$, $m' = m$, $t' = t$, but ε' can be different from ε.

---

[10] See, for instance, for the finite ordered case, H. W. Kuhn, Contributions to the Theory of Games, Vol. II, p. 209.

[11] See C. Berge, "Sur une généralisation du théorème de Zermelo-Von Neumann," C. R. Ac. des Sc., tome 241, p. 455, 1955. See also, "Topological Games," Colloque of the Int. Conf. of Math., Vol. II, Amsterdam 1954.

The aim of the Agent is to reach a position x (with the move to the Agent), ξ lying in the set

$$K = \{x = (s, t, m, \xi)/s = s_0;\ t \leq t_0;\ m \geq 0\} .$$

Let us define the set K' such that

$$K' = \{x = (s, t, m, \xi)/s \neq s_0;\ t > t_0 \text{ or } m < 0\} .$$

The game terminates if the position lies in $K \cup K'$ (with the move to the Agent).

The payoff is the characteristic function of the set K, i.e., the function $f(x) = 0$ if $x \notin K$, $f(x) = 1$ if $x \in K$.

The game is a one person topological game, since the space X is the topological product $S \times T \times M \times [0, 1]$ of four topological spaces, and Γ_1, Γ_2 are continuous multivalued functions on X to X; moreover K is a closed set in X, and the payoff function f(x) is, therefore, an upper semi-continuous function on $K \cup K'$. It is not assumed that the alternative sets be finite, or that there should be a uniquely defined path leading from a given position x_1 to another x_2; but it is assumed that if a position x occurs at a time t_0, then the same position x cannot occur at a time $t_1 \neq t_0$. Moreover, if the game is a mission, the lengths of the plays are bounded.

Under these assumptions, it seems important to study whether there exists a good strategy σ_1.

Let $z(x, i, \sigma_1, \sigma_0)$ be the last position of the game when the initial position is x, the Agent adopting the strategy σ_1, chance adopting σ_0, and i denoting the player which has the move at the beginning of the game. If $p(\sigma_0)$ is the probability attached to the strategy σ_0, the expected payoff, for the Agent adopting the strategy σ_1, is

$$F(x, i, \sigma_1) = \sum_{\sigma_0} p(\sigma_0)\, f[z(x, i, \sigma_1, \sigma_0)] .$$

A multivalued function $\Gamma(x)$ from a topological space X to a topological space Y is said to be upper semi-continuous if

1° $\Gamma(x)$ is a compact set for all x in X;

2° $\Gamma^+(\Omega) = \{x/x \in X;\ \Gamma(x) \neq \phi$: the empty set; $\Gamma(x) \neq \phi;\ \Gamma(x) \subset \Omega\}$ is open for all open sets Ω in Y.

THEOREM: Given a mission for which X is a topological space, f(x) is an upper semi-continuous function on $K \cup K'$, $\Gamma_1(x)$ and $\Gamma_0(x)$ are upper semi-continuous multivalued functions on X, $\{p(y)/y \in \Gamma_0(x)\}$ is a continuous probability distribution, then there exists a good strategy σ_1 for the Agent.

PROOF: Consider the sets

$$C_0 = K \cup K'$$
$$C_1 = \Gamma_0^+ (C_0)$$
$$C_2 = \Gamma_1^+ (C_1)$$
$$C_3 = \Gamma_0^+ (C_2)$$
$$\cdots\cdots\cdots$$

C_{2k} is the set of all the positions at the time t_0-k (with the move to the Agent). It is possible to assume that $K \cup K'$ is an open set (if it is not so, one can use the method of the "topological sum"). The sets C_{2k} are open, disjoints, and $X = \bigcup_{k=0}^{m} C_{2k}$.

We shall define σ_1 successively on $C_0, C_2, C_4, \ldots$; σ_1 will be defined on C_0 by $\sigma_1(x) = x$; if σ_1 is defined on C_{2k-2}, for every x in C_{2k}, one will take $\sigma_1(x) = y$ such that $y \in \Gamma_1(x)$, and

$$F(y, 0, \sigma_1) = \max_{z \in \Gamma_1(x)} F(z, 0, \sigma_1) ;$$

y exists if $F(z, i, \sigma_1)$ is an upper semi-continuous function for $i = 0, 1$; it is therefore necessary to prove that it is so. Let us assume that it is true on C_{k-1} and let us show that it is still true on C_k.

One has:

$$F(x_0, 1, \sigma_1) = \max_{z \in \Gamma_1(x_0)} F(z, 0, \sigma_1)$$

$$F(x_0, 0, \sigma_1) = \sum_{z \in \Gamma_0(x_0)} p(z)\, F(z, 1, \sigma_1).$$

Consider the open set

$$\Omega = \{z / z \in C_{k-1} ;\ f(z, 0, \sigma_1) < M_0 + \epsilon\}$$

where:

$$M_0 = \max_{x \in \Gamma_1(x_0)} f(x, 0, \sigma_1) .$$

There exists a neighbourhood $V(x_0)$ of x_0 such that

$$x \in V(x_0) \text{ implies } \Gamma_1(x) \subset \Omega$$

Thus,

$$x \in V(x_0) \text{ implies } F(x, 1, \sigma_1) < M_0 + \epsilon ;$$

$F(x, 1, \sigma_1)$ is therefore an upper semi-continuous function on C_k (and, of course, $F(x, 0, \sigma_1)$ is also an upper semi-continuous function on C_k). In the case of the problem of the boat, the Agent has therefore a definite good strategy, according to the theorem above, if the set of all the prohibited places is open in S. A construction of the good strategy σ_1 is given in the proof of the theorem.

* * *

VARIOUS OPTIMIZERS FOR SINGLE-STAGE PRODUCTION

Wayne E. Smith[1]

The following note develops theorems which enable one to schedule operations on a single-stage production system so as to minimize certain functions or optimizers. It is shown that the maximum tardiness of job completion is minimized by ordering the jobs according to the date the job is due to be completed. The sum of times until each job is completed may be minimized under various conditions. If all jobs can be completed by their due dates, a weighted sum of completion times can be minimized, subject to all jobs being completed by their due dates. This is done by scheduling last the jobs with the largest ratio of processing time to weighting factor among those jobs which will be completed on time if scheduled last; the next to last job is scheduled by applying the same rule to the remaining jobs, and so forth, until the complete ordering is obtained.

In the process of production scheduling there are so many factors that must be taken into consideration that the determination of a best schedule is impossible without a very precise meaning of the word "best." We shall discuss some fairly simple criteria which might be used as a means of judging whether or not a schedule is good. We shall also show how the choice of an optimizer will affect the characteristics of best schedules, in certain special cases.

We shall be concerned with single-stage production and with optimizers such that there is never any advantage in delaying the completion of any operation, unless the completion of at least one other operation is moved forward. It is not difficult to see that a schedule is then completely determined by the ordering of the objects to be processed; that is, an optimal solution subject to a given ordering is obtained by performing each operation as soon as possible, consistent with the ordering.

SECTION I

In most of the scheduling with which we will be concerned, we establish sufficient conditions that an ordering will be a "best" ordering in some sense of the word. We wish now to establish a sufficient condition that we shall use repeatedly.

We assume that the desirability of any schedule can be measured, and that this measurement can be expressed as a function f over the set of all schedules. Since the schedule is assumed to be completely determined by the ordering, then f can be considered as a function of the ordering or as a function over the set of permutations of the n objects to be processed. We wish to find a permutation π, where π is the order $(\pi_1, \pi_2, \ldots, \pi_n)$, which minimizes the value of f.

---

[1]Prepared while the author was under contract to the Office of Naval Research, Logistics Branch, and employed on the Management Sciences Research Project, University of California, Los Angeles. The author wishes to thank Dr. J. R. Jackson and A. J. Rowe for their helpful suggestions and comments.

THEOREM: A sufficient condition that $f(\bar{\pi}) \leq f(\pi)$ for all permutations π of the n objects is that:

(A) there is a real-valued function g of ordered pairs of elements such that if π is any permutation and π' the permutation obtained from π by the interchange of the i-th and (i+1)-st elements, then $f(\pi) \leq f(\pi')$ if $g(\pi_i, \pi_{i+1}) \leq g(\pi_{i+1}, \pi_i)$ and

(B) $\bar{\pi}$ is such that the i-th object precedes the j-th object if $g(i,j) < g(j,i)$.

PROOF: Given that the permutation $\bar{\pi}$ satisfies condition (B) we wish to show that any $\pi \neq \bar{\pi}$ can be altered by a series of interchanges of pairs of adjacent elements in such a manner that the value of the function decreases or remains the same at every interchange. Since $\bar{\pi} \neq \pi$, there is an element π_i of π such that π_i is immediately preceded by an item which follows π_i in the order $\bar{\pi}$. (For example, π_i may be the element appearing as $\bar{\pi}_j$, where the j-th position is the first position at which π and $\bar{\pi}$ do not agree.) Therefore, $g(\pi_i, \pi_{i-1}) \leq g(\pi_{i-1}, \pi_i)$ so we may interchange π_i and π_{i-1} in the permutation π, obtaining a new permutation π' with $f(\pi') \leq f(\pi)$. Also π' has one fewer pairs of elements which disagree in ordering with those of $\bar{\pi}$. One may proceed inductively, decreasing the number of pairs of elements which disagree in ordering with those of $\bar{\pi}$ and at each step decreasing the value of the function or leaving the value of the function the same.

The proof becomes more simple if g is a function which relates the elements in a transitive manner, that is if $g(i,j) \leq g(j,i)$ and $g(j,k) \leq g(k,j)$ together imply that $g(i,k) \leq g(k,i)$. However, this is not necessary, as can be seen from the example in S. M. Johnson's paper [1].

It may be possible to simplify the function g so that it is a function of only one of the elements. In this case rule (B) becomes: $\bar{\pi}$ is such that the i-th object precedes the j-th object if $g(i) < g(j)$. This does lead to a transitive relationship and is the type of relationship that is found in most of the following examples.

We have not stated any method of obtaining the function g. In the following examples the function g is derived from the function f. Upon interchanging the i-th and (i+1)-st elements of a permutation, the value of f changes by a certain amount. If this change can be attributed to quantities attached to just the i-th and (i+1)-st items, then one may be able to eliminate constant values from both values of the function f, obtaining $f(\pi) = A + B\, g(\pi_i, \pi_{i+1})$ and $f(\pi') = A + B\, g(\pi_{i+1}, \pi_i)$, where π' differs from π only by the interchange of the i-th and (i+1)-st elements. It may then be possible to simplify this by finding a function h such that $g(i,j) < g(j,i)$ if $h(i) < h(j)$.

In the specific examples that follow we shall assume we are given the operation times (which shall include the setup times except in Section VIII) for a finite number of jobs. Also, the optimizers with which we are concerned are such that the deletion of all idle times cannot make a schedule worse so we shall assume that no idle time occurs.

SECTION II

We now assume that each operation or job has a due date and that we wish to minimize the maximum tardiness. This type of optimizer might be used if the output of the machine is used in an assembly line production system where the delay on the assembly line is the maximum of the individual delays.

We let T_i be the operation time of the i-th object and D_i be the due date of the i-th object. If the objects are processed in numerical order the tardiness of the i-th object is:

$\delta_i = T_1 + T_2 \ldots + T_i - D_i$. Let $\Delta = \max \{\delta_i\}$ over $i = 1, 2, \ldots, n$. Our problem is to find an ordering which minimizes Δ.

THEOREM:[2] The order $\{1,2,3,\ldots,n\}$ is an order which minimizes Δ if $D_1 \leq D_2 \leq D_3 \leq \ldots \leq D_n$.

PROOF: Let π be the order 1, 2, 3, . . . , n, and let π' be the same order except with i and i+1 interchanged. Thus $T_j = T'_j$ and $D_j = D'_j$ for $j \neq i$ or $i+1$; $T_i = T'_{i+1}$, $T_{i+1} = T'_i$, $D_i = D'_{i+1}$ and $D_{i+1} = D'_i$. Also, $\delta_j = \delta'_j$ for $j \neq i$ or i+1. Thus $\Delta \leq \Delta'$ if

$$\text{Max } (\delta_i, \delta_{i+1}) \leq \text{Max } (\delta'_i, \delta'_{i+1}) .$$

This is by definition equivalent to:

$$\text{Max } \left(\sum_{j=1}^{i} T_j - D_i, \sum_{j=1}^{i+1} T_j - D_{i+1}\right) \leq \text{Max } \left(\sum_{j=1}^{i} T'_j - D'_i, \sum_{j=1}^{i+1} T'_j - D'_{i+1}\right) .$$

Replacing the T'_i by their equal T_i values we have:

$$\text{Max } \left(\sum_{j=1}^{i} T_j - D_i, \sum_{j=1}^{i+1} T_j - D_{i+1}\right) \leq \text{Max } \left(\sum_{j=1}^{i-1} T_j + T_{i+1} - D_{i+1}, \sum_{j=1}^{i+1} T_j - D_i\right) .$$

Subtracting from each side $\sum_{j=1}^{i+1} T_j$

$$\text{Max } (-T_{i+1} - D_i, -D_{i+1}) \leq \text{Max } (-T_i - D_{i+1}, -D_i)$$

or

$$-\text{Min } (T_{i+1} + D_i, D_{i+1}) \leq -\text{Min } (T_i + D_{i+1}, D_i)$$

or

$$\text{Min } (T_i + D_{i+1}, D_i) \leq \text{Min } (T_{i+1} + D_i, D_{i+1}) .$$

Since all T_i are positive, if $D_i \leq D_{i+1}$ this condition is met, the D_i being the smallest of the four terms concerned. Therefore, if $D_i \leq D_{i+1}$ we have that $\Delta \leq \Delta'$, and the order with i preceding i+1 is as good or better than the order with them interchanged. We have, then, by the theorem of Section I, that the maximum tardiness can be minimized if the jobs are arranged according to their due dates. Thus, if it is possible to complete all the jobs by their due dates, this can be done by ordering them according to due date.

---

[2] J. R. Jackson [2] has generalized this theorem to take into consideration restrictions on the time at which the processing of each item can start.

SECTION III

In scheduling for single-stage production, there is a waiting time, a time that each job must wait until it can be processed. Thus, if the objects are processed in numerical order, the waiting time for the j-th object is $\sum_{i=1}^{j-1} T_i$. One optimizer that may be considered is the sum of all waiting times, $\sum_{j=1}^{n} \sum_{i=1}^{j-1} T_i$. The minimum of this expression occurs at the same time as the minimum of the expression $\sum_{j=1}^{n} \sum_{i=1}^{j} T_i$ occurs, since the two expressions differ by the constant, $\sum_{i=1}^{n} T_i$. We shall use the latter formulation in this paper.

If the objects are processed in numerical order, the completion time of the j-th object is $C_j = \sum_{i=1}^{j} T_i$. We see that the above problem is the same as the problem of finding an order which minimizes the sum of completion times, $\sum_{i=1}^{n} C_i$.

THEOREM: $\sum_{i=1}^{n} C_i$ is minimized over all orderings of the n objects if the objects are arranged in increasing order of values of T_i.

PROOF: $\sum_{i=1}^{n} C_i = \sum_{i=1}^{n} \sum_{j=1}^{i} T_j = \sum_{j=1}^{n} (n+1-j)\, T_j$. According to Hardy, Littlewood and Polya (Inequalities, p. 261), such a sum is least when the series are monotonic in opposite senses. Therefore, the optimizer will have its minimum value if $T_1 \le T_2 \le T_3 \le \ldots \le T_n$.

SECTION IV

We may now generalize the theorem of Section III. Since some jobs may be considered more important than others, we may weight the jobs accordingly and consider a weighted sum of completion times. Weighting job i by an amount $a_i \ge 0$, the optimizer is then of the form $\sum_{j=1}^{n} a_j C_j$.

THEOREM: The sum $\sum_{j=1}^{n} a_j C_j$ is minimized over all orderings of the n objects if the objects are arranged so the T_i/a_i form a non-decreasing sequence.

PROOF: Let π be the order 1,2,3,...,n and let π' be the same order except with i and i+1 interchanged. Thus $T_j = T'_j$ and $a_j = a'_j$ for $j \neq i$ or i+1; $T_i = T'_{i+1}$, $T_{i+1} = T'_i$, $a_i = a'_{i+1}$, $a_{i+1} = a'_i$. Also $C_j = C'_j$ for $j \neq i$ or i+1. Now, the order π is as good as or better than π' if

$$\sum_{j=1}^{n} a_j C_j \le \sum_{j=1}^{n} a'_j C'_j .$$

By the above equalities, this is true if

$$a_i C_i + a_{i+1} C_{i+1} \leq a'_i C'_i + a'_{i+1} C'_{i+1}$$

or

$$a_i \left(\sum_{j=1}^{i} T_j \right) + a_{i+1} \left(\sum_{j=1}^{i+1} T_j \right) \leq a_{i+1} \left(\sum_{j=1}^{i-1} T_j + T_{i+1} \right) + a_i \left(\sum_{j=1}^{i+1} T_j \right).$$

Subtracting $a_i \left(\sum_{j=1}^{i} T_j \right) + a_{i+1} \left(\sum_{j=1}^{i-1} T_j + T_{i+1} \right)$ from both sides this becomes

(1) $$a_{i+1} T_i \leq a_i T_{i+1}.$$

If a_i and a_{i+1} are not zero, we may divide by their product obtaining

(2) $$T_i/a_i \leq T_{i+1}/a_{i+1}.$$

Defining $T/0$ to be infinity, condition (1) and condition (2) are equivalent whether or not any a's are zero.

Therefore if condition (2) is satisfied, the order with i preceding i+1 is as good as or better than the order with them interchanged. Thus, by the theorem of Section I, the minimum value of the optimizer is obtained if the objects are arranged in increasing order on T_i/a_i.

SECTION V

We now propose to use the same optimizer as in Section III, the sum of times until each job is completed, but add the restriction that each job must be completed by its given due date, D_i.

THEOREM: If all jobs, 1,2,3,..., n, can be completed by their respective due dates,[3] then there is an ordering of the jobs with the k-th job last which minimizes $\sum_{i=1}^{n} C_i$, subject to completion of all jobs by their respective due dates if, and only if, the k-th job has the properties:

(i) $D_k \geq \sum_{i=1}^{n} T_i$ and

(ii) $T_k \geq T_i$ for all i for which $D_i \geq \sum_{i=1}^{n} T_i$.

PROOF: Let π be the ordering 1,2,3,..., n, and let it be so chosen that with the objects in this order all jobs can be completed by their due dates. Let I_n be the set of indices for which $D_i \geq \sum_{i=1}^{n} T_i$. Since the ordering π is such that all jobs are completed by their due dates, we have

[3] We stated in Section II that if a set of jobs can be completed by their due dates, this can be done by ordering according to their due dates.

that n is in the set I_n. Suppose that k is the index of the object with the largest value of T_i for i in the set I_n. (If this is not unique, we choose one arbitrarily from those which are the largest.) Now let π' be the ordering obtained from π by interchanging the k-th and n-th terms. Thus $T_i = T'_i$ and $D_i = D'_i$ for $i \neq k$ or n; $T_k = T'_n$, $T_n = T'_k$, $D_k = D'_n$ and $D_n = D'_k$. Now for $j < k$, $C'_j = \sum_{i=1}^{j} T'_i = \sum_{i=1}^{j} T_i = C_j$, and since $T_k \geq T_n$ we have for $k \leq j < n$, $C'_j = \sum_{i=1}^{j} T'_i = \sum_{i=1}^{j} T_i + T_n - T_k = C_j + T_n - T_k \leq C_j$. Also, $C'_n = \sum_{i=1}^{n} T'_i = \sum_{i=1}^{n} T_i = C_n$. Therefore, taking the sums of C's, we have that $\sum_{i=1}^{n} C'_i \leq \sum_{i=1}^{n} C_i$. Also, all jobs are completed by their respective due dates under the new ordering, since for $i \neq k$ or n we have $C'_i \leq C_i \leq D_i = D'_i$; $C'_k \leq \sum_{i=1}^{n} T_i \leq D_n = D'_k$ and $C'_n = \sum_{i=1}^{n} T_i \leq D_k = D'_n$.

In the particular case that π is an ordering which minimizes $\sum_{i=1}^{n} C_i$, we have found an ordering π' which also minimizes the optimizer. Thus, if the k-th object has properties (i) and (ii) there is an ordering of the jobs with the k-th job last, which minimizes the optimizer subject to completion of all jobs by their respective due dates.

If the original ordering π has a last item which does not have property (ii), then $T_n < T_k$, and in the above steps for $k \leq j < n$, $C'_j < C_j$ so that $\sum_{i=1}^{n} C'_i < \sum_{i=1}^{n} C_i$. Thus π cannot be an ordering which minimizes the optimizer. The ordering π is not even admissible without property (i).

In order to obtain a complete ordering which minimizes $\sum_{i=1}^{n} C_i$, we can determine an item to be scheduled last, one which has properties (i) and (ii). Once having determined the last item, C_n is fixed, and we have the problem of ordering n-1 jobs so as to minimize $\sum_{i=1}^{n-1} C_i$, subject to the completion of all jobs by their respective due dates. Applying the theorem again, we can determine the last of these n-1 jobs. Continuing in this manner, we can find a complete ordering which minimizes the value of the optimizer.

We note that if the objects are arranged in increasing order on T_i and still meet their due dates, then one cannot decrease the value of the optimizer by rearrangement, that is, the value of the optimizer for this order is a lower bound on its value no matter what the side restrictions are.

On the other hand, if the objects are arranged in increasing order on D_i, and this order is such that $D_{j-1} < \sum_{i=1}^{j} T_i \leq D_j$ for all j (taking $D_0 = 0$), then the given order is the only arrangement by which all jobs can be completed on time.

SECTION VI

The results of Sections IV and V can be combined to give a method of ordering so as to minimize the weighted sum $\sum_{i=1}^{n} a_i C_i$ subject to the restriction that every job must be completed

by its due date, that is, $C_i \leq D_i$. If all jobs can be completed by their due dates, an order which minimizes the weighted sum may be obtained. This order has as its last job one with the largest value of T_i/a_i from among those with due date as large as the total processing time of all jobs. Proceeding as before, one can determine the last of the remaining n-1 jobs to be processed and continue in like manner until the complete ordering is determined.

SECTION VII

Other optimizers that have been considered are:

$$\text{Total tardiness} = \sum_{i=1}^{n} \max (0, C_i - D_i)$$

$$\text{Weighted tardiness} = \sum_{i=1}^{n} a_i [\max (0, C_i - D_i)] .$$

Thus far we can give solutions only in special cases. For example, if all the T_i are the same, the total tardiness is minimized if the objects are arranged in order of due dates. If the objects are arranged in order of due dates and then have only one tardy item, this order minimizes total tardiness since it has minimized the maximum tardiness. In minimizing total tardiness, a rule for the ordering of two adjacent operations is as follows: If the first of the i-th and (i+1)-st operations start at time t_0, and if $\max \{D_i, t_0 + T_i\} \leq \max \{D_{i+1}, t_0 + T_{i+1}\}$, then the order $\{i,i+1\}$ gives no greater total tardiness than the order $\{i+1,i\}$. Since this relation depends on t_0, one cannot in general expect to reach the optimal solution by only a series of interchanges of adjacent items.

SECTION VIII

Thus far we have considered only the case where the setup times could be included as part of the operation times. Under these assumptions, the over-all completion time of n given jobs is fixed and is just the sum of the T_i's, regardless of order. Now let us see what type of problem we have if we wish to minimize the over-all completion time of all jobs where the setup times are taken into consideration.

Since the sums of the actual operation times is constant, and the over-all completion time is determined by the sum of operation times and setup times, our problem is equivalent to minimizing the sum of the setup times. This is a non-symmetric "traveling salesman" type of problem. To be more precise, the amount of setup times required for the j-th job, if it follows the i-th job, is taken to be the "distance" from i to j (not necessarily the same as from j to i). We then have the problem of finding a cycle or "tour" starting with 0 (the original setup), passing through each object once, and only once, and returning to 0, which minimizes the distance traveled. (If one does not care in what form the machine is set up at the end of the operations, then the distance from j to 0 would all be taken as zero.) Since such problems have been discussed elsewhere, we shall not consider this problem further [3, 4].

SECTION IX

The optimizers for which an optimum ordering has been found in this paper are such that the solution can be obtained from any order (satisfying the given restrictions) by a series of interchanges of adjacent operations, improving (or not making worse) the value of the optimizer at each step. Thus there are no "local" optima with respect to permutations of adjacent elements. We have shown that any optimizer which satisfies the conditions of the theorem of Section I also has this property. Although it should not be too difficult to test to see whether specific optimizers satisfy these conditions, it would be desirable to characterize more completely those optimizers for which no "local" optima exist. Further research is planned in this direction.

Another difficult problem to be considered, of course, is that of finding how to determine an order which minimizes the value of similar functions when several machines or stages of production are used.

REFERENCES

[1] Johnson, S. M., "Optimal Two- and Three-Stage Production Schedules with Setup Times Included," Naval Research Logistics Quarterly, Vol. 1, No. 1, 1954, pp. 61-68.

[2] Jackson, J. R., "Scheduling a Production Line to Minimize Maximum Tardiness," Research Report No. 43, 36 pp, Management Sciences Research Project, University of California, Los Angeles.

[3] Dantzig, G., Fulkerson, R., and Johnson, S., "Solution of a Large Scale Traveling Salesman Problem," Journal of the Operations Research Society of America, Vol. 2, No. 4, November 1954.

[4] Kuhn, H. W., "The Traveling Salesman Problem," Proceedings of the Sixth Symposium in Applied Mathematics of the American Mathematical Society, McGraw-Hill Book Co.

* * *

NOTES ON THE THEORY OF DYNAMIC PROGRAMMING
IV - MAXIMIZATION OVER DISCRETE SETS

Richard Bellman
The RAND Corporation

The theory of dynamic programming is applied to a class of problems involving maximization over discrete sets. The solution is made to depend on the solution of a class of functional equations.

INTRODUCTION

A problem of frequent occurrence is that of determining the maximum of a function $F(x_1, x_2, \ldots, x_N)$, subject to the constraints

$$\text{(1)} \qquad \begin{aligned} &\text{(a)} \quad G_i(x_1, x_2, \ldots, x_N) \le c_i, \; i = 1, 2, \ldots, K \\ &\text{(b)} \quad x_i \in S_i, \end{aligned}$$

where S_i is a discrete, usually finite, set. The most important case is that where each S_i is a finite set of integers, and an interesting sub-case is that where $x_i = 0$ or 1.

A particular class of problems of this type concerns the maximization of

$$\text{(2)} \qquad F(x) = \sum_{i=1}^{N} F_i(x_i),$$

over the set of x_i constrained by the relations

$$\text{(3)} \qquad \begin{aligned} &\text{(a)} \quad \sum_{j=1}^{N} G_{ij}(x_j) \le c_i, \; i = 1, 2, \ldots, K \\ &\text{(b)} \quad x_i \in S_i, \; i = 1, 2, \ldots, K, \end{aligned}$$

with $G_{ij}(x_j) \ge 0$ for $x_j \in S_j$, $G_{ij}(0) = 0$, and $0 \in S_i$ for all i.

Even in the case where the F_i and G_{ij} are linear functions of the x_i, this problem at the moment escapes any of the standard computational algorithms of linear programming, such as the simplex method of G. Dantzig.

We shall show that this problem may be treated by means of the functional equation technique of the theory of dynamic programming [1], and that this technique yields a very simple computational solution whenever the number of constraints is small.

We shall also indicate the application of the method to a problem involving mutually exclusive activities. Here we have an additional constraints of the form

$$\text{(4)} \qquad x_i x_{i+1} = 0, \; i = 1, 2, \ldots, N - 1.$$

FUNCTIONAL EQUATION

Let us define the sequence of functions

$$f_N(c_1, c_2, \ldots, c_K) = \max_{\{x\}} \sum_{i=1}^{N} F_i(x_i), \tag{5}$$

where the x_i are subject to the constraints of (3). Then

$$f_1(c_1, c_2, \ldots, c_K) = \max_{x_1} F_1(x_1) \tag{6}$$

where

$$\text{(7)} \quad \begin{aligned} &\text{(a)} \quad G_{11}(x_1) \le c_1, \ldots, G_{K1}(x_1) \le c_K, \\ &\text{(b)} \quad x_1 \in S_1. \end{aligned}$$

Applying the principle of optimality, we obtain the recurrence relation

$$f_N(c_1, c_2, \ldots, c_K) = \max_{x_N} [F_N(x_N) + f_{N-1}(c_1 - G_{1N}(x_N), \ldots, c_K - G_{KN}(x_N))], \tag{8}$$

where

$$\text{(9)} \quad \begin{aligned} &\text{(a)} \quad G_{1N}(x_N) \le c_1, \ldots, G_{KN}(x_N) \le c_K. \\ &\text{(b)} \quad x_N \in S_N. \end{aligned}$$

EXAMPLE

Consider the problem of determining the maximum of $L_N(x) = \sum_{i=1}^{N} a_i x_i$ subject to the constraints

$$\text{(10)} \quad \begin{aligned} &\text{(a)} \quad \sum_{i=1}^{N} b_i x_i \le c, \\ &\text{(b)} \quad x_i = 0 \text{ or } 1, \end{aligned}$$

where $a_i, b_i > 0$.

Here

$$\text{(11)} \quad \begin{aligned} f_1(c) &= a_1, \quad c \ge b_1, \\ &= 0, \quad c < b_1, \end{aligned}$$

and

$$\text{(12)} \quad \begin{aligned} f_N(c) &= \max_{x_N = 0,1} [a_N x_N + f_{N-1}(c - b_N x_N)], \quad c \ge b_N \\ &= f_{N-1}(c), \quad c < b_N. \end{aligned}$$

DISCUSSION

The functions $f_N(c)$ may now be computed with ease on either a digital or hand computer, depending upon the size of the system, starting with the known value $f_1(c)$.

To give an example, suppose that $N = 50$ and $c = 200$, with the a_i, b_i integers ranging between 1 and 10. The naive approach involves the testing of 2^{50} sets of values, i.e., all possible combinations of accept or reject. Since $2^{50} \cong 10^{50(.30)} = 10^{4.5}$, this is a considerable task. Conventional linear programming techniques fail because of the restriction that the x_i be integral. For the case where $N = 50$, a round-off of the linear programming solution may cause considerable error.

Using the above method, we must compute 50 functions $\{f_N(c)\}$, each containing 200 entries, $c = 1, 2, 3, \cdots$. If the a_i and b_i are irrational, we may have to refine the c-grid in order not to introduce round-off errors of importance. An important point to note is that doubling the size of N will double the computational time, which is to say that the time required for computing the solution in this fashion is proportional to N, rather than dependent upon N in some exponential fashion, as in ordinary search methods.

In return for the labor expended in computing the sequence $f_N(c)$, one has all the advantages of a "sensitivity analysis." It is easy to trace the influence of c and N upon the maximum value and the behavior of the maximizing $x_N = x_N(c)$.

Let us now discuss in more detail the remark we made in the introduction stating that this technique is, at the present time, restricted to problems involving a small number of constraints.[1]

Consider a cargo-handling problem in which we have a number of items possessing values v_i, weights w_i, and sizes s_i. We wish to maximize the value of the cargo carried, subject to weight restriction w and a volume restriction s.

The mathematical problem is that of maximizing

$$L(x) = \sum_{i=1}^{N} x_i v_i, \tag{13}$$

subject to the restrictions

$$\text{(a)} \quad \sum_{i=1}^{N} x_i w_i \le w,$$

$$\text{(b)} \quad \sum_{i=1}^{N} x_i s_i \le s \tag{14}$$

$$\text{(c)} \quad x_i = 0, 1, 2, \ldots$$

Defining

$$f_N(w, s) = \text{Max } L(x), \tag{15}$$

we readily obtain

[1] We have recently developed a new method, based upon a combination of the Lagrange multiplier method of classical variational theory and the functional equation method of dynamic programming, which greatly enlarges the scope of the methods presented here. A discussion of these methods will appear subsequently.

$$f_1(w, s) = v_1 \operatorname{Min}\left(\left[\frac{w}{w_1}\right], \left[\frac{s}{s_1}\right] \right),$$

(16)

$$f_k(w, s) = \operatorname*{Max}_R \left[v_k x_k + f_{k-1}(w - x_k w_k, s - x_k s_k) \right],$$

where R is the set

$$x_k = 0, 1, 2, \ldots, \operatorname{Min}\left(\left[\frac{w}{w_k}\right], \left[\frac{s}{s_k}\right] \right). \tag{17}$$

Taking the parameters w_i, s_i, and v_i to be integers, we will, in general, be required to N functions of two variables, tabulated at the points of a grid $w = 0, 1, 2, \ldots, \overline{W}$, $s = 0, 1, 2, \ldots, S$. If W and S are of the order of magnitude of 100, this requires 10^4 values. This is still within the capability of modern machines.

It is clear, however, that one additional constraint of the same type puts us in the 10^6 range. This exceeds the capability of any present day machine.

If, on the other hand, there are a large number of constraints, each with a small range, then the method is useful.

EXAMPLE—MUTUALLY EXCLUSIVE ACTIVITIES

Consider the problem of the preceding section under the additional constraint

$$x_i x_{i+1} = 0, \quad i = 1, 2, \ldots, N - 1 \tag{18}$$

Define the sequence of functions

$$f_N(c, b) = \operatorname*{Max}_{\{x\}} \sum_{i=1}^{N} a_i x_i \tag{19}$$

where the x_i are subject to the constraints

$$\text{(a)} \quad x_n \cdot b = 0, \quad b = 0 \text{ or } 1$$

(20)

$$\text{(b)} \quad \sum_{i=1}^{N} b_i x_i \le c.$$

Then we have the recurrence relation

$$f_N(c, b) = \operatorname*{Max}_{x_N=0,1} \left[a_N x_N + f_{N-1}(c - b_N x_N; x_N) \right]. \tag{21}$$

To determine the solution we must compute the double sequence $\{f_N(c, 0), f_N(c, 1)\}$.

REFERENCE

[1] Bellman, R., "The Theory of Dynamic Programming," Bull. Amer. Math. Soc., Vol. 60 (1954), pp. 503-516.

* * *

ATTRITION GAMES

J. R. Isbell[1]
Aberdeen Proving Ground
and
W. H. Marlow
The George Washington University
Logistics Research Project

This paper is a summary report of investigations of C. Tompkins, the authors, and others concerning rather general mathematical models of military situations. The basic assumptions on the mechanism of attrition are given by certain differential equations modified and generalized from the classical Lanchester equations by R. N. Snow, C. Tompkins and the authors. Part I deals with the attrition equations and is thus a study of a family of stochastic processes. It is also the study of combat wherein strategy in a game theoretic sense does not enter; the combat is described by physical constants expressed as probabilities--no provision is made for rational choices by the players.

Parts II and III deal with games, i.e. minimax problems, which isolate certain aspects of the problem of military (strategic, tactical, or logistical) decision. In Part II the decision means the commitment of a part of the available force to battle, retaining the remainder as a reserve. In Part III the forces are in the field, and the decision means a distribution of the fire of each type of unit among the several possible types of targets. Part IV is devoted to the problem of estimating intrinsic values of weapons. Results from Parts I and II, along with related work by R. H. Brown, are employed to motivate a linear estimate for weapons of one type against one enemy type. This estimate is then generalized for the case of several weapon types.

More precise details are given in the introductions to the individual parts of the paper. It will be clear to the reader that this paper, while contributing somewhat to the study of attrition and military decision, presents several unsolved problems. Of particular note are difficult computational problems on a theoretical level.

PART I - ATTRITION PROCESSES

In this Part we are concerned with bivariate stochastic processes based upon probabilities of survival of individual units. We summarize work of various authors [9, 11, 12] with some added results and comments. The central problem is to determine $P(x, \xi; r, \rho; t)$ for all arguments non-negative and x, ξ, r, ρ integral; P is the probability that at time t, (x, ξ) units survive of (r, ρ) which began at $t = 0$. With the exception of I.4, we consider homogeneous attrition: individual units available to a player are identical but not necessarily like units available to his opponent. In I.1 we consider a probabilistic case: P is positive for $t > 0$, $0 \leq x \leq r$ and $0 \leq \xi \leq \rho$. We give a partial solution in closed form for a special case (I.1.6) and indicate that numerical computation of the general case appears promising. We next introduce a truncated process, Section I.2, in which we ignore the behavior of the process in time in favor of the direction of travel of the system. This process yields to analysis and is justified somewhat by

[1]Research supported (in part) by the Office of Naval Research. This article was presented to the American Mathematical Society September 1955, December 1955, and February 1956.

conclusions we draw from subsequent games. Section I.3 treats deterministic processes of which the Lanchester equations are standard examples. Finally, the non-homogeneous attrition of Section I.4 is based on Ref. [11]; we limit our consideration to aspects required for Part III.

1. PROBABILISTIC ATTRITION

We consider a homogeneous attrition process described by $P(x, \xi; r, \rho; t)$ as introduced above in which $f(x, \xi)dt$ ($\phi(x, \xi)dt$) is the probability of loss of one of the first (second) player's units if a total of x oppose a total of ξ for time dt. If the probability of simultaneous loss of two or more units is assumed to be negligible, the (forward) transition system is

$$\frac{dP(x, \xi; \ldots)}{dt} = f(x+1, \xi)\, P(x+1, \xi; \ldots) + \phi(x, \xi+1)\, P(x, \xi+1; \ldots)$$

$$- [f(x, \xi) + \phi(x, \xi)]\, P(x, \xi; \ldots); \quad x = 0, 1, \ldots, r,$$

$$\xi = 0, 1, \ldots, \rho; \tag{1}$$

$$P(x, \xi; r, \rho; 0) = \delta_r^x \delta_\rho^\xi;$$

$$P(x, \xi; r, \rho; t) = 0 \text{ for all } t \text{ if } x > r \text{ or } \xi > \rho.$$

We are concerned with the case,

$$f(x, \xi) = ax + \beta\xi, \quad f(o, \xi) = 0;$$

$$\phi(x, \xi) = \alpha\xi + bx, \quad \phi(x, o) = 0. \tag{2}$$

Assumption (2) is roughly that each x-unit has chance a of death not caused by the enemy and b of destroying an enemy unit while α and β have similar meanings for a ξ-unit. For the case (2) there is clearly a unique solution to (1) (cf. [4, pp. 409, 416]), but the reader should notice that the discontinuities in (2) cause the differential equations for $x > 0 = \xi$ and $\xi > 0 = x$, not to be special cases of the equations for $x\xi > 0$. The reader can verify that the exact integration of (1) for literal a, b, α, β is straightforward through, say, $r-2 \leqq x \leqq r, \rho-2 \leqq \xi \leqq \rho$, but that it readily becomes impractical for smaller x and ξ (for large r and ρ).

The backward (retrospective) equation [5, p. 384] corresponding to (1) and (2) is

$$\frac{dP(\ldots; r, \rho; t)}{dt} = (ar + \beta\rho)\, P(\ldots; r-1, \rho; t) + (\alpha\rho + br)\, P(\ldots; r, \rho-1; t),$$

$$- [(a+b)r + (\alpha+\beta)\rho]\, P(\ldots; r, \rho; t), \tag{3}$$

where in case $r = 0$ or $\rho = 0$ we may preserve the formula by setting $P(\ldots; -1, \rho; t)$ and $P(\ldots; r, -1, t)$, respectively, equal to $P(\ldots; o, \rho, t)$ and $P(\ldots; r, o; t)$. The easiest proof of (3) is to write P' instead of $P(x, \xi; r, \rho; t)$ in the right-hand member of the forward equation (1). Then P' is trivially an analytic solution of the forward equation and, by computation, it satisfies the initial condition of (1).

The backward equation is convenient because it is satisfied by any function $F(t, r, \rho) = \sum_{x, \xi} c(x, \xi)\, P(x, \xi; r, \rho; t)$; e.g., the expected value of x at time t, given by $c(x, \xi) = x$,

satisfies (3) and therefore can be computed by recursion on r and ρ without reference to individual values of P. The backward equation is solvable in a special case which unfortunately is not very important; however, the solution and argument leading to it give an indication of the difficulties of the general case.

Consider first the time-free probability $F(x, \xi; r, \rho)$ that there is some time $t \geq 0$ at which the numbers of survivors are exactly x, ξ. Of course, $F(r, \rho; r, \rho) = 1$; and F satisfies the dual recursion relations

$$(4)\quad F(x,\xi;r,\rho) = \frac{a(x+1)+\beta\xi}{(a+b)(x+1)+(\alpha+\beta)\xi} F(x+1,\xi;r,\rho) + \frac{\alpha(\xi+1)+bx}{(a+b)x+(\alpha+\beta)(\xi+1)} F(x,\xi+1;r,\rho)$$

$$(5)\quad F(x,\xi;r,\rho) = \frac{ar+\beta\rho}{(a+b)r+(\alpha+\beta)\rho} F(x,\xi;r-1,\rho) + \frac{\alpha\rho+br}{(a+b)r+(\alpha+\beta)\rho} F(x,\xi;r,\rho-1)\,.$$

One can derive (4) and (5) from previous formulas, but the reader might as well regard (5) as a definition. Adding $F(x, \xi; x, \xi) = 1$ and $F(x, \xi; r, \rho) = 0$ if $x > r$ or $\xi > \rho$, one has a complete definition of F. It is the correct definition only if $x > 0$ and $\xi > 0$; hence our following argument applies only to this case.

Supposing that $a + b = \alpha + \beta$, one can solve (3) with the initial conditions of (1), obtaining

$$(6)\qquad P(x,\xi;r,\rho;t) = F(x,\xi;r,\rho)\binom{r+\rho}{x+\xi}(e^{(a+b)t}-1)^{r+\rho-x-\xi}e^{-(a+b)(r+\rho)t}\,.$$

Given the formula (6), one easily verifies it; we omit the computation. We were led to (6) by argument patterned after C. A. Truesdell [13] as follows. Replace P with the function $Q = e^{(a+b)(r+\rho)t}$; replace t with the variable $z = \dfrac{e^{(a+b)t}}{a+b}$. Now, Q and z are so constructed that (3) transforms into the simple pattern equation

$$(7)\qquad \frac{dQ(x,\xi;r,\rho;z)}{dz} = (ar+\beta\rho)\,Q(\ldots;r-1,\rho;z) + (br+\alpha\rho)\,Q(\ldots;r,\rho-1;z)\,,$$

eliminating the undifferenced term on the right side of the equality. In Truesdell's one-dimensional problem this can always be done; he gives explicit rules. In our problem it is this step which requires the hypothesis $a + b = \alpha + \beta$. (While we obviously cannot prove that the problem is "harder" otherwise, we have a more complicated, quite different, derivation of (6) which requires the same special hypothesis.)

Now, by iterating (7), the derivatives of $Q(x, \xi; r, \rho; z)$ are expressed in terms of certain functional values $Q(x, \xi; r', \rho', z)$, where r' and ρ' are smaller than r and ρ. If the unknown function Q is analytic in z, then it can be written as a Taylor series. Expanding about the value of z which corresponds to $t = 0$, we find from the initial conditions that all terms of the series vanish except the $(r+\rho-x-\xi)$-th. Hence Q is known in terms of z, and therefore P is known in terms of t.

The iteration mentioned at the beginning of the preceding paragraph leads to complicated coefficients. One can in principle keep track of the coefficients, $\phi(x, \xi; r, \rho)$, which turn out to be $P(x, \xi; r, \rho)\binom{r+\rho}{x+\xi}$ as in (6). We found it easier to observe that if $x' + \xi' = x + \xi$, then (since

$a + b = \alpha + \beta$) the ratio of $P(x', \xi'; r, \rho; z)$ to $P(x, \xi; r, \rho, z)$ is independent of time and is the same as the ratio of the corresponding ϕ's or F's. Hence, it suffices to find $\sum_{x+\xi=k} \phi(x, \xi; r, \rho) = G(k; r, \rho)$. (The corresponding sum of F's is unity since the total number of survivors is k once.) Now the probability that the number was k at time t and diminished before t + dt, may be written

$$(a + b)k\,(e^{(a+b)t} - 1)^{r+\rho+k} e^{-(a+b)(r+\rho)t}\, G(k; r, \rho)\, dt\,;$$

the integral of this expression from t = 0 to infinity is unity (interpreted as a probability), and by calculation this is $G(k; r, \rho)/\binom{r+\rho}{k}$. Hence we obtain (6).

Numerically, the attrition process (1) and (2) can be solved by recursion on a digital computer with standard techniques. One fixes a, b, α, β to a factor of proportionality and then tabulates the P's for desired values of t. Such tabulated probability distributions seem to be what is required for the games to be presented in Sections II.1 and II.2.

One criticism of (2) as a description of attrition rates (in addition to its apparent difficulty of analysis) concerns firepower: the firepower of each unit contributes to the chance of destruction of each enemy unit. We remark first that for solution on a computer it is practical to have f and ϕ non-linear; consideration of particular alternates to (2) is left to the reader who has a special physical situation to approximate. What does seem to be to the point is the obvious answer that we do not wish to rule out the consideration of a battle as a collection of independent events: each event or engagement can be described by a system such as (1).

Suppose that the physical situation dictated that never more than two units would fire on the same enemy unit (cf. [12]). An engagement would then have (r, ρ) equal to one of (2, 1), (1, 2), (1, 1), (1, 0) or (0, 1) cases for which (1) is easy to solve. For any (r, ρ), one requires that the player with numerical superiority can arrange the (unique) most favorable (to him) set of engagements. Then the probabilities P for the battle (r, ρ) are computed by recursion: for any (r, ρ) there is a "nearest" (r', ρ'), $r' \leq r$, $\rho' \leq \rho$ such that the battle (r, ρ) consists of the battle (r', ρ') plus one additional engagement. The average number of survivors from the battle is of course the sum of the average survivors from engagements; the probability of winning (destroying all enemies) is a simple product; other such expressions are readily computed.

2. A TRUNCATED PROCESS

In this section we remove the continuous variable t in favor of the discrete $x + \xi$, the total number of units surviving. The discrete steps in the battle correspond to the successive losses of individual units. As before, we rule out simultaneous destruction of two or more units and start with (1.1,2). This time we suppose $r = x$ and $\rho = \xi$ and specify that at most one unit is lost. We obtain

$$P(x, \xi; \ldots) = \exp\{-[(a+b)x + (\alpha+\beta)\xi]\,t\}\,,$$

(1)
$$P(x-1, \xi; \ldots) = \frac{ax + \beta\xi}{(a+b)x + (\alpha+\beta)\xi}\,[1 - P(x, \xi; \ldots)]\,,$$

$$P(x, \xi-1; \ldots) = \frac{\alpha\xi + bx}{(a+b)x + (\alpha+\beta)\xi}\,[1 - P(x, \xi; \ldots)]\,.$$

From (1) we find easily

$$\lim_{t\to\infty} P(x-1, \xi; \ldots) = \frac{ax+\beta\xi}{(a+b)x+(\alpha+\beta)\xi} = P^-(x,\xi),$$

(2)

$$\lim_{t\to\infty} P(x, \xi-1; \ldots) = 1 - P^-(x,\xi) = P^+(x,\xi).$$

We label these probabilities P^+ for the favorable event from the first players point of view and P^- for the unfavorable. For any x, ξ, r, ρ, t in the process (1.1, 2), P^+ is the probability that, given x and ξ survive at t, the next unit lost will be a ξ-unit.

We imagine that this time-free process starts with r vs. ρ and is continued until one force is annihilated. If the end results (x, o), (o, ξ) are evaluated by a function u, the values of u can be computed recursively by the formula

(3) $$u(x,\xi) = P^-(x,\xi)\,u(x-1,\xi) + P^+(x,\xi)\,u(x,\xi-1);$$

i.e., the value at (x, ξ) is this convex combination of the values at the possible successors of (x, ξ). For example, if $v(x, o) = 1$ for $x > 0$, $v(o, \xi) = 0$ ($v(0, 0)$ need not be defined), then v is the probability that the first player will win. Again, if $v(x, 0) = x$, $v(0, \xi) = 0$, v computed by (3) is the unconditional number of x-survivors from a fight to the finish combat.

It should be clear that the truncated process of this section can be used if one assumes new strategic decisions can be made whenever a unit is destroyed. That is, the application is not solely to fights to the finish; we develop this in II.3.

3. DETERMINISTIC PROCESSES

In this section we are concerned with the average progress of a battle: we ask for $x' = x(t)$ and $\xi' = \xi(t)$ such that $P(x', \xi'; r, \rho, t) = 1$.

The classical Lanchester equations [9, Chapter 5], in case there is no replacement of losses, are

$$\frac{dx}{dt} = -ax - \beta\xi,$$

(1)

$$\frac{d\xi}{dt} = -\alpha\xi - bx,$$

$$x(0) = r, \quad \xi(0) = \rho.$$

The solutions to equations (1) are given in [9, p. 72]. If $a = \alpha = 0$, the flow lines are hyperbolas with asymptotes given by $x^2 b = \xi^2 \beta$. Thus one speaks of the Law of n^2; the effectiveness of a force is measured by multiplying its firepower by the square of its size. In any case, there is one straight flow line (in positive x and ξ), given by

(2) $$\xi = \frac{\alpha - a + \sqrt{(\alpha-a)^2 + 4b\beta}}{2\beta}\,x.$$

If $b\beta > a\alpha$, the other flow lines bend away from the line (2) with increasing t; thus an initial advantage is increased as the fight continues. If $a\alpha > b\beta$, the flow lines bend toward (2). If $a\alpha = b\beta$, the flow lines are straight lines parallel to (2). In some degenerate cases one player always wins, but in general there is a critical ratio, given by the slope of (2) such that the winner can be determined by evaluating each x-unit at this price in terms of ξ-units; the winner is the force of greater value. We have cited the Law of n^2 above for the case $a = \alpha = 0$, similar descriptions of the combat for other cases are easy. We return to the problem of estimating intrinsic values of units in Part IV.

It is of interest to compare (1) with equations satisfied by the average number of survivors in (1.1, 2). Upon multiplication by x and summing and similarly for ξ, these latter equations yield (cf. [11, p. 24]):

$$
\begin{aligned}
\frac{d\bar{x}}{dt} &= -a\bar{x} - \beta\bar{\xi} + \beta \sum_{\xi} \xi P(0, \xi; \ldots), \\
\frac{d\bar{\xi}}{dt} &= -\alpha\bar{\xi} - b\bar{x} + b \sum_{x} x P(x, 0; \ldots), \qquad (3) \\
\bar{x}(0) &= r\,; \quad \bar{\xi}(0) = \rho\,.
\end{aligned}
$$

Direct calculation with (3) is impossible unless the summation terms are known. In view of our I.1 this would seem to indicate that $\bar{x}$ and $\bar{\xi}$ could then be computed from their definition. We observed above that $\bar{x}$ and $\bar{\xi}$ satisfy the backward equation (1.3) so that such a recursive calculation appears to be indicated for values t such that (1) is an inadequate approximation to (3).

In summary, the deterministic processes which we have noted are: the Lanchester equations (1), the exact equations (3), the readily solved expressions for average survivors noted in the last paragraph of section I.1 and equations (2.3) for survivors in the truncated process.

4. NON-HOMOGENEOUS ATTRITION

In [11, p. 22], equations generalizing (1.1) are given for combat of a force of m unit types of strength r_i in the i-th type for $i = 1, 2, \ldots, m$ against a force similarly described by $\rho_1, \ldots, \rho_n$. We shall not write these equations since our primary interest in probabilistic non-homogeneous attrition is a generalization of the truncated process of Section I.2. As the analogue of (1.2) we assume for r_i respectively ρ_j non-zero:

$$
\begin{aligned}
f_i(r_1, \ldots, r_m\,;\ \rho_1, \ldots, \rho_n) &= a_i r_i + \sum_j \rho_j \beta_{ji} \pi_{ji}\,, \qquad (1) \\
\phi_j(r_1, \ldots, r_m\,;\ \rho_1, \ldots, \rho_n) &= \alpha_j \rho_j + \sum_i r_i b_{ij} p_{ij}\,;
\end{aligned}
$$

e.g., the probability of loss of one unit of type i in combat of time dt is $f_i \cdot dt$ if $r_i \neq 0$ and is zero if $r_i = 0$.

As usual, we have introduced variables in Latin-Greek pairs and it will suffice to explain one from each pair. Here a_i represents a rate of exponential decay of i-type friendly forces and b_{ij} is a measure of effectiveness of fire of i-type friendly units against j-type enemies. As a new concept (and one to be exploited in Part III), we have introduced the strategic variables p_{ij}, the proportional part of the total effort of units of type i (for player I) which is directed

against units of type j. Here we are using strategic in a game theoretic sense; in military language the p_{ij}'s are tactical variables.

We restrict ourselves to a study of the optimal choice of the p's and π's in Part III wherein (cf. (2.1)) it suffices to look only as far as the first casualty. If the symbol d_i refers to the possible event that the first unit killed is a friendly unit of type i, we require $Pr(d_i)$ and the analogous probability $Pr(\delta_j)$. As in (2.2) we obtain

$$Pr(d_i) = \frac{a_i r_i + \sum_j \rho_j \beta_{ji} \pi_{ji}}{\sum_k a_k r_k + \sum_j \alpha_j \rho_j + \sum_{j,k} (r_k b_{kj} p_{kj} + \rho_j \beta_{jk} \pi_{jk})} . \tag{2}$$

Finally, for the sake of completeness, we remark that the corresponding generalization of the deterministic equations (3.1) is as follows for $i = 1, \ldots, m$ respectively $j = 1, \ldots, n$:

$$\frac{dx_i}{dt} = -f_i(x_1, \ldots, x_m ; \xi_1, \ldots, \xi_n) , \quad x_i(0) = r_i ,$$

$$\frac{d\xi_j}{dt} = -\phi_j(x_1, \ldots, x_m ; \xi_1, \ldots, \xi_n) , \quad \xi_j(0) = \rho_j . \tag{3}$$

For fixed p's and π's, these equations are treated in [11, pp. 12-17]. Their solution is naturally much more complicated than the solution of the Lanchester equations; however, as we have remarked, our interest is in the selection of the p's and π's in the probabilistic case for which (4.2) suffices.

PART II - HOTSPOT PROBLEMS

We begin with a summary of C. Tompkins' paper on Hotspot games. These games are probabilistic in that they are based on attrition such as described in Section I.1; their calculation presents several difficulties. In Section II.2 we present a related matrix game (for which there are several efficient known methods of solution). There is a heuristic connection between these matrix games and Hotspot games but, more importantly, whenever one has a solution in pure strategies then so does the other, with the same good strategies and value. Finally, in Section II.3 we introduce a simplified game inspired by the Hotspot model. Instead of having new assignments of forces to battle at stated intervals, we permit a new assignment for both players whenever a unit on either side is destroyed. This simplified game has a solution in pure strategies; hence neither player would wish to change his assignment at any other time than those prescribed. The best strategy for either player in any case is to assign all, none, or exactly one of his units to combat. The games of Section II.3 are thus disposed of; they are used in Parts III and IV as a basis for generalizations toward more difficult problems.

1. THE HOTSPOT GAME

In this section we summarize [12] in which C. Tompkins introduced a zero-sum two-person game concerning assignment of forces to a disputed area, "Hotspot." The game is not a matrix game; however, it is reducible to a one-move game.

391368 O - 56 - 6

For the k-th move, (r_k, ρ_k) units from total resources (m_k, μ_k) are assigned from safe havens to the Hotspot area. The result of a move is possible loss of units by one or both players; it is supposed that a homogeneous attrition function $P(x, \xi; r, \rho, t')$ of Part I applies where t' is constant and does not vary between moves. No units can be introduced after play has started. The game terminates at move k if $r_k \cdot \rho_k = 0$, in which case payoff to Player I for Hotspot is the quantity F, $-\Phi$, or 0 according as $r_k \neq 0$, $\rho_k \neq 0$, or $r_k = \rho_k = 0$, respectively. In addition, there is the payoff $-(m_1 - m_k) + \sigma(\mu_1 - \mu_k)$, i.e., Player I pays 1 for each of his units lost and receives σ for each of Player II's units destroyed. Thus, we adopt the usual convention that Player I is the maximizing player.

Tompkins proved that the game can be considered to terminate in a finite number of moves and thus established the existence of a unique value, $v(m, \mu)$, to the many-move game [10]. In order to provide for more efficient calculation than the constructive method of his existence proof, the next step was to reduce the game to a single move. Since $v(m, \mu)$ does exist uniquely, one can write the following expectations of Player I for the first move where (r, ρ) are assigned to Hotspot from total resources (m, μ) and each player plays optimally thereafter:

$$E(r, \rho; m, \mu) = A(m, \mu; r, \rho) + (B(r, \rho) + 1)\, v(m, \mu) . \tag{1}$$

Here, A is the expectation from all possible outcomes of the move which result in loss of at least one unit; or precisely,

$$\begin{aligned} A(0, 0; m, \mu) &= 0 \\ A(r, 0; m, \mu) &= F - \sum_x (r - x)\, P(x, 0; r, 0; t') , \quad r \neq 0 ; \\ A(0, \rho; m, \mu) &= -\Phi + \sigma \sum_\xi (\rho - \xi)\, (P(0, \xi; 0, \rho, t') , \quad \rho \neq 0 ; \end{aligned} \tag{2}$$

$$A(r, \rho; m, \mu) = \sum_{x, \xi} [v(m - r + x, \mu - \rho + \xi) - (r - x) + \sigma(\rho - \xi)] \cdot P(x, \xi; r, \rho; t'), \quad r \neq 0, \rho \neq 0 .$$

where the summation extends over $0 \leq x \leq r$, $0 \leq \xi \leq \rho$ less the point (r, ρ). For manipulative convenience,

$$\begin{aligned} B(r, \rho) &= P(r, \rho; r, \rho; t') - 1 , \quad r \cdot \rho \neq 0 ; \\ &= -1, \text{ otherwise. .} \end{aligned} \tag{3}$$

It should be noted first that A is independent of m and μ for $r \cdot \rho = 0$. Indeed, for these cases $A = E$, so that games where one or both players have no resources can be disposed of immediately. Suppose next that an inductive hypothesis is made: $v(r, \rho)$ is known for $r \leq m$ and $\rho \leq \mu$ with the sole exception of $v = v(m, \mu)$. Then, all quantities in (2) and (3) are known and a straightforward procedure would be to define v to be the value of the game with payoffs (1), or equivalently, the root of

$$w(v) = \text{Value } (A + Bv) = 0 . \tag{4}$$

Tompkins established the validity of this definition and gave an iterative method of calculating v. The existence of a unique root in (4) has also been proved by von Neumann [14], Loomis [8], and

Dines [3] under the hypothesis that the elements of the matrix B are negative. For further related results, see Bellman [1].

The Hotspot game is thus reduced to a one-move game whose calculation is somewhat different from standard matrix games: its value and optimal strategies result from making the matrix game in (4) have zero value. In order to find $v(m, \mu)$, $(m+1)(\mu+1)$ successive systems (4) must be solved.

2. A HOTSPOT VARIANT

As we have just seen, solving the original Hotspot game means determining a number v and optimal strategies such that the game G_1 with payoff matrix $(a_{ij} + b_{ij} v)$ has value zero. There is a solution in pure strategies (i, j) if, and only if, $v = -a_{ij}/b_{ij}$, which suggests the game G_2 with payoff matrix $(a_{ij}/-b_{ij})$. In fact,

(1) If each $b_{ij} < 0$, G_1 has a solution in pure strategies (r, s) and $v = a_{rs}/-b_{rs}$ if, and only if, (r, s) is a solution of G_2.

PROOF: If G_1 has such a solution, say v^*, the remaining payoff elements, $a_{ij} + b_{ij} v^*$, in row r (column s) will be non-negative (non-positive). Since the slope of the line $E = a_{ij} + b_{ij} v$ is negative, the v-intercepts satisfy $a_{rs}/-b_{rs} = \min_j a_{rj}/-b_{rj} = \max_i a_{is}/-b_{is}$, i.e., (r, s) is a solution of G_2. The argument is clearly reversible to obtain the converse.

The following example shows that neither the game values nor even the pure strategies entering mixed strategy solutions need agree.

$$A = \begin{pmatrix} 5 & 1 \\ 1 & 4-\epsilon \\ 1 & 1 \end{pmatrix}, \qquad -B = \begin{pmatrix} 1 & 3 \\ 5 & 2-\epsilon \\ 3 & \frac{1}{2} \end{pmatrix}.$$

In G_1, $v = 1$ and Player II plays rows one and two. On the other hand, for small ϵ, value $(G_2) > 1$ and rows one and three are played. We have examples where the game values do not agree under mixed strategy solution for A and B given by (1.2, 3). We have not been able to find interesting restrictive classes of A's and B's permitting comparison of strategies for G_1 and G_2 which enter into optimal mixed strategy solutions.

We conclude this section with a formal connection between the games G_1 and G_2. Returning to Section II.1, proceed as before through equations (1.3). Then assume that at least one unit will be lost in every move with $r \cdot \rho \neq 0$ and, furthermore, $P(x, \xi; r, \rho; t')$ is to be replaced by $Q(x, \xi; r, \rho; t')$;

$$Q(r, \rho; r, \rho; t') = 0$$

$$Q(x, \xi; r, \rho; t') = \frac{P(x, \xi; r, \rho; t')}{1 - P(r, \rho; r, \rho; t')}.$$

By this process of discounting the chance that there will be no units lost and then renormalizing the P's, we are again led to a matrix game with payoffs $A(r, \rho; m, \mu)/-B(r, \rho)$.

Unless each $v(m, \mu)$ is determined by a game with saddle point, this new procedure may lead to different v's so that the A's which have just been written may not coincide with those for the original Hotspot game. It is not clear how close the connection between these two procedures actually is. Numerical experience has indicated that the connection may be stronger than is indicated by (2.1).

3. A SIMPLIFIED GAME

The game of this section is based upon the truncated attrition process of Section I.2. Two commanders, disposing respectively of r, ρ units, dispute the possession of an objective. If they assign x respectively ξ units (retaining reserves $r - x, \rho - \xi$), and if $x \cdot \xi > 0$, there ensues an attrition process governed by (I.2.1). If $x \cdot \xi = 0$ the battle is over and any non-zero force on the battlefield has gained the objective. Whenever a unit on either side is lost, both commanders are free to make new assignments of their remaining forces.

One has a game as soon as values of the terminal states are specified. For each r and ρ, one has a game with an $(r+1)$ by $(\rho+1)$ matrix as soon as five values are known: $v^+(r, \rho)$, the value to the first player if the other assigns no force, so that the first player wins the objective; $v^-(r, \rho)$, the value of the opposite situation; $v(r, \rho)$, the value if both players give up the fight; and the game values $w(r-1, \rho)$, $w(r, \rho-1)$. There are numerous inequalities one would expect the v's to satisfy. Specifically, we assume $v^- \leqq v \leqq v^+$.

The payoff to the strategies (x, ξ) in the game (r, ρ) is (I.2.2):

$$H(x, \xi; r, \rho) = P^+(x, \xi)\, W(r, \rho-1) + P^-(x, \xi)\, W(r-1, \rho). \tag{1}$$

The formula (1) naturally applies only for $x \cdot \xi \neq 0$. One has also $H(x, 0; r, 0) = v^+(r, 0)$ and the symmetric formula for $(0, \rho)$; the case $r = \rho = 0$ need not be considered as it cannot arise from any other case.

The $(r+1)$ by $(\rho+1)$ matrix H always has a saddle point. The main reason is that $P^+(x+1, \xi) \geqq P^+(x, \xi) \geqq P^+(x, \xi+1)$ if, and only if, $b\beta \geq a\alpha$ and all these inequalities reverse together. Then, if $b\beta > a\alpha$, the strategy r (and ρ) dominates all others, except possibly 0, and we reduce immediately to the matrix

$$\begin{pmatrix} v^*(r, \rho) & v^+(r, \rho) \\ v^-(r, \rho) & v(r, \rho) \end{pmatrix} \tag{2}$$

where $v^*(r, \rho) = H(r, \rho)$. But $v^- \leqq v \leqq v^+$; hence the question is where v^* lies.

For the case $b\beta \geq a\alpha$,

$$\begin{aligned} &\text{if } v^* > v^+, \ v^+ \text{ is a saddle point of (2);} \\ &\text{if } v^- \leq v^* \leq v^+, \ v^* \text{ is a saddle point;} \\ &\text{if } v^* < v^-, \ v^- \text{ is a saddle point.} \end{aligned} \tag{3}$$

Similarly,

$$\text{if } b\beta \leq a\alpha, \text{ substitute } v^* = H(1, 1) \text{ for } v^* \text{ in (3); the result is true.} \tag{3'}$$

The conclusion (3') rests essentially on the special assumption $H(x, 0)$ = constant for $x > 0$, $H(0, \xi)$ = constant for $\xi > 0$. But (3) would hold on the reasonable weaker hypothesis $H(x, 0)$ monotone in x, $H(0, \xi)$ anti-monotone. The degeneration of these games confirms the qualitative observation that one should usually commit all available force to action (no other, simultaneous action being in view); for usually $b\beta \geq a\alpha$. (From (3) and (3'), if $b\beta = a\alpha$, then all non-zero assignments are equivalent.)

The entire course of a fight governed by (3) is simply as follows: both forces are entirely committed and continue fighting until one gives up the fight (possibly being annihilated). Under (3'), on the other hand, each force sends in one unit; when the "champion" of a force is lost, it is replaced, until finally one force abandons the field. Then any modification such as cost of moving or advantage of temporary possession of the field will incline to make the familiar principle given by (3) even more generally applicable.

PART III. PROGRAMMING GAMES

In this part we deal with a model involving the isolated strategic problem of distribution of fire over enemy targets of several types. The attrition is described by functions introduced in Part I and the model of the game has a design closely similar to the last model considered in Part II. It turns out that the present model can be solved by comparatively simple computation, for interesting mathematical reasons which we set forth in some detail. We conclude with an examination of "extensive programming games," finding that combination of decisions on assignment of force with decisions on distribution of fire presents serious difficulties, which we do not overcome.

1. FIRE PROGRAMMING GAMES

Formally, a fire programming game is described by a payoff function

$$H(P, \Pi) = \sum_i Pr(d_i)\, w(d_i) + \sum_j Pr(\delta_j)\, w(\delta_j) . \tag{1}$$

The strategic variables, P, Π, represent distributions of the fire of several weapons over several enemy targets. In (1) we have suppressed explicit reference to the forces engaged in the game. There are m types of units engaged with n types of enemy units; there are r_i friendly units of the i-th type ($i = 1, 2, \ldots, m$) and ρ_j units of the j-th enemy type. Then P is an $m \times n$ matrix of real numbers p_{ij}, subject to the constraints

$$\begin{aligned} &p_{ij} \geq 0 \quad \text{for all } i, j ; \\ &\sum_j p_{ij} = 1 \quad \text{for each } i . \end{aligned} \tag{2}$$

We interpret p_{ij} as that fraction of the total firepower of the i-type friendly units which is directed at enemies of type j. Similarly the enemy has a firing strategy $\Pi = (\pi_{ji})$, subject to corresponding constraints. Now H is the value of the situation to our side; thus P is chosen to maximize $H(P, \Pi)$, and Π to minimize it.

The symbol d_i refers to the possible event that the first unit killed is a friendly unit of type i. We suppose that the resulting situation has been evaluated already, and has value $w(d_i)$.

Unless each $v(m, \mu)$ is determined by a game with saddle point, this new procedure may lead to different v's so that the A's which have just been written may not coincide with those for the original Hotspot game. It is not clear how close the connection between these two procedures actually is. Numerical experience has indicated that the connection may be stronger than is indicated by (2.1).

3. A SIMPLIFIED GAME

The game of this section is based upon the truncated attrition process of Section I.2. Two commanders, disposing respectively of r, ρ units, dispute the possession of an objective. If they assign x respectively ξ units (retaining reserves $r - x, \rho - \xi$), and if $x \cdot \xi > 0$, there ensues an attrition process governed by (I.2.1). If $x \cdot \xi = 0$ the battle is over and any non-zero force on the battlefield has gained the objective. Whenever a unit on either side is lost, both commanders are free to make new assignments of their remaining forces.

One has a game as soon as values of the terminal states are specified. For each r and ρ, one has a game with an $(r+1)$ by $(\rho+1)$ matrix as soon as five values are known: $v^+(r, \rho)$, the value to the first player if the other assigns no force, so that the first player wins the objective; $v^-(r, \rho)$, the value of the opposite situation; $v(r, \rho)$, the value if both players give up the fight; and the game values $w(r-1, \rho)$, $w(r, \rho-1)$. There are numerous inequalities one would expect the v's to satisfy. Specifically, we assume $v^- \leqq v \leqq v^+$.

The payoff to the strategies (x, ξ) in the game (r, ρ) is (I.2.2):

$$H(x, \xi; r, \rho) = P^+(x, \xi)\, W(r, \rho-1) + P^-(x, \xi)\, W(r-1, \rho). \tag{1}$$

The formula (1) naturally applies only for $x \cdot \xi \neq 0$. One has also $H(x, 0; r, 0) = v^+(r, 0)$ and the symmetric formula for $(0, \rho)$; the case $r = \rho = 0$ need not be considered as it cannot arise from any other case.

The $(r+1)$ by $(\rho+1)$ matrix H always has a saddle point. The main reason is that $P^+(x+1, \xi) \geqq P^+(x, \xi) \geqq P^+(x, \xi+1)$ if, and only if, $b\beta \geq a\alpha$ and all these inequalities reverse together. Then, if $b\beta > a\alpha$, the strategy r (and ρ) dominates all others, except possibly 0, and we reduce immediately to the matrix

$$\begin{pmatrix} v^*(r, \rho) & v^+(r, \rho) \\ v^-(r, \rho) & v(r, \rho) \end{pmatrix} \tag{2}$$

where $v^*(r, \rho) = H(r, \rho)$. But $v^- \leqq v \leqq v^+$; hence the question is where v^* lies.

For the case $b\beta \geq a\alpha$,

(3)
- if $v^* > v^+$, v^+ is a saddle point of (2);
- if $v^- \leq v^* \leq v^+$, v^* is a saddle point;
- if $v^* < v^-$, v^- is a saddle point.

Similarly,

(3′) if $b\beta \leq a\alpha$, substitute $v^* = H(1, 1)$ for v^* in (3); the result is true.

The conclusion (3') rests essentially on the special assumption $H(x, 0)$ = constant for $x > 0$, $H(0, \xi)$ = constant for $\xi > 0$. But (3) would hold on the reasonable weaker hypothesis $H(x, 0)$ monotone in x, $H(0, \xi)$ anti-monotone. The degeneration of these games confirms the qualitative observation that one should usually commit all available force to action (no other, simultaneous action being in view); for usually $b\beta \geq a\alpha$. (From (3) and (3'), if $b\beta = a\alpha$, then all non-zero assignments are equivalent.)

The entire course of a fight governed by (3) is simply as follows: both forces are entirely committed and continue fighting until one gives up the fight (possibly being annihilated). Under (3'), on the other hand, each force sends in one unit; when the "champion" of a force is lost, it is replaced, until finally one force abandons the field. Then any modification such as cost of moving or advantage of temporary possession of the field will incline to make the familiar principle given by (3) even more generally applicable.

PART III. PROGRAMMING GAMES

In this part we deal with a model involving the isolated strategic problem of distribution of fire over enemy targets of several types. The attrition is described by functions introduced in Part I and the model of the game has a design closely similar to the last model considered in Part II. It turns out that the present model can be solved by comparatively simple computation, for interesting mathematical reasons which we set forth in some detail. We conclude with an examination of "extensive programming games," finding that combination of decisions on assignment of force with decisions on distribution of fire presents serious difficulties, which we do not overcome.

1. FIRE PROGRAMMING GAMES

Formally, a fire programming game is described by a payoff function

$$H(P, \Pi) = \sum_i P\,r(d_i)\; w(d_i) + \sum_j P\,r(\delta_j)\, w(\delta_j)\,. \tag{1}$$

The strategic variables, P, Π, represent distributions of the fire of several weapons over several enemy targets. In (1) we have suppressed explicit reference to the forces engaged in the game. There are m types of units engaged with n types of enemy units; there are r_i friendly units of the i-th type ($i = 1, 2, \ldots, m$) and ρ_j units of the j-th enemy type. Then P is an $m \times n$ matrix of real numbers p_{ij}, subject to the constraints

$$\begin{aligned} &p_{ij} \geq 0 \quad \text{for all } i, j\,; \\ &\sum_j p_{ij} = 1 \quad \text{for each } i\,. \end{aligned} \tag{2}$$

We interpret p_{ij} as that fraction of the total firepower of the i-type friendly units which is directed at enemies of type j. Similarly the enemy has a firing strategy $\Pi = (\pi_{ji})$, subject to corresponding constraints. Now H is the value of the situation to our side; thus P is chosen to maximize $H(P, \Pi)$, and Π to minimize it.

The symbol d_i refers to the possible event that the first unit killed is a friendly unit of type i. We suppose that the resulting situation has been evaluated already, and has value $w(d_i)$.

Thus we implicitly suppose (a) that we have listed all possible developments and, in particular, forces will not be added to nor subtracted from the battle before a kill occurs, and further (b) that a kill will be recognized immediately by both commanders, or at least that the value of the result does not depend on the strategies P and Π (as it would if there were a time lag in changing strategies). Note that we had corresponding assumptions in Section II.3; (a) was supported by the result, but the argument for (b) is merely simplicity.

Interchanging Greek and Latin (excepting the subscripts) turns d_i into δ_j, the corresponding event that the first kill is a j-type enemy. The probabilities $\mathrm{P}\,r(d_i)$ depend on P and Π by the formula (I.4.2) which we relabel

$$\mathrm{P}\,r(d_i) = \frac{a_i r_i + \sum_j \rho_j \beta_{ji} \pi_{ji}}{\sum_k a_k r_k + \sum_j \alpha_j \rho_j + \sum_{j,k} (r_k b_{kj} p_{kj} + \rho_j \beta_{jk} \pi_{jk})} \tag{3}$$

With (1), the Greek-Latin pair (2), and the Greek-Latin pair (3), we have completed the definition of a fire programming game. The assumed manner of attrition, described extensively by the coefficients (I.4.1), is essentially the simplest form which has any pretensions of describing actual combat. The move rule of the game is also the simplest available; formally, we have only one move, but in effect, we assume that a move, i.e., a change of firing strategy, can be made instantaneously as soon as there is any reason to move. Finally, we have limited the commanders to making decisions on one aspect of tactics, namely the apportionment of fire among targets.

Solution of the present problem turns out to be remarkably easy, for reasons which cannot be expected to apply when any of our simplifying assumptions are removed nor even if the model is set up a little differently without being more intrinsically complex. Hence we wish to emphasize at this point that the problem we are faced with is the determination of optimal strategies P and Π varying with time and with the observed strategy of the opponent. Such problems are quite difficult; but in this model we shall find optimal constant strategies.

2. COMPUTATION METHODS

Before solving a fire programming game we shall consider the simpler problem of linear fractional programming, the maximization of a linear fractional function under linear constraints. One is given a convex polyhedron X in real n-space, and a function

$$f(x) = \frac{t \cdot x + h}{u \cdot x + k}; \tag{1}$$

one must find $\max_{x \in X} f(x)$. Now for linear functions f, numerous effective methods are known; this is the problem of linear programming. We shall describe a method for linear fractional programming which involves solution of linear programming problems. First, however, we shall describe the problem in geometric language.

The simplicity of (1) lies in the fact that it is essentially a two-dimension transformation, in the following sense. The denominator vanishes on some hyperplane D, and the numerator vanishes on some hyperplane N. Usually (Case I) D and N are not parallel, but meet in an (n-2)-dimensional subspace L of the given n-dimensional space E. The function f is undefined on L; it is constant on each subspace parallel to L. Thus we may as well project E

along L upon a plane, in which D and N become lines meeting in a point L. The function f is constant on lines passing through L; on any circle around L, f goes through all real values monotonically, twice, jumping from $+\infty$ to $-\infty$ each time one crosses D. Thus we may clarify our description as follows: f is a one-dimensional transformation, having singularities which appear both as infinities (on D) and as essential singularities (on L), but we can apply first a two-dimensional transformation with no singularities. All the complications appear in the plane, and even there the situation is rather simple; the reader familiar with homogeneous coordinates for the projective line will recognize them here.

Turning to the polyhedron X, transformed to a polygon by the above projection, we see that it cannot meet D if a maximum is to exist. Thus it must lie on one side of D. Here f increases monotonically as x moves in a direction which we may call clockwise; we wish to find the farthest ray through L, going clockwise, which touches X. For such to exist, one must have further conditions on X, which we omit to specify precisely; it suffices if X is bounded.

Let us write the problem

$$\max_{x \in X} \frac{t \cdot x + h}{u \cdot x + k} = v . \tag{2}$$

The denominator has constant sign, which we may take positive. Then multiplying out preserves inequalities, and we have

$$\max_{x \in X} (t - vu) \cdot x = vk - h . \tag{3}$$

That is, the ray $f(x) = v$, which is the farthest clockwise in X, is also the farthest in the family of all lines parallel to it, which still touches X. Unfortunately we do not know this parallel family to begin with. However, we can approximate it. The method is

Method: To find $\max_{x \in X} \frac{t \cdot x + h}{u \cdot x + k}$ in a bounded polyhedron X,

(1) Choose x^1 in X, and inductively

(2) Compute $w^n = \frac{t \cdot x^n + h}{u \cdot x^n + k}$;

(3) Solve $\max_{x \in X} (t - w^n u) \cdot x = w^n k - h + d$.

(i) In (3), $d \geq 0$. If $d = 0$ then $w^n = v$ and x^n is optimal. Otherwise

(4) Let x^{n+1} be optimal in (3) and repeat (2).

(ii) The values w_i increase monotonically to v and the iteration terminates, i.e., $w^n = v$ for some n.

The formal proof consists of a few computations. If $d < 0$ in (3) then $w^n > v$, which is impossible by definition of w^n. The values w^n increase monotonically because we are steadily approximating toward the extreme ray $f(x) = v$. The iteration terminates because the solution of each linear problem (3) is a vertex of X, and there are only finitely many of these.

We have omitted Case II; D and N are parallel. In this case $t = au$ in (1) for some constant a. The method above applies, and $w^2 = v$. Geometrically, there are no singularities in this case, and f is a projective transformation onto the projective line. The interested reader may verify this for himself. We shall also omit working out an analogous method for certain minimax problems more general than fire programming games, namely

$$\max_{(a,b)\in X_1} \min_{(c,d)\in X_2} \frac{a+c}{b+d}, \tag{4}$$

where (a, b) and (c, d) are points in the plane., Fire programming games have this form, where a, b, c and d are linear functions of several real variables which happen to be subject to the very convenient constraints (1.2).

Let us write $F = \sum_i f_i$ and $\Phi = \sum_j \phi_j$. From (I.4.1), (III.1.1), and (III.1.3) we have

$$H(P, \Pi) = \frac{\sum_j f_j(P)\, w(\delta_j) + \sum_i \phi_i(\Pi)\, w(d_i)}{F(P) + \Phi(\Pi)} \tag{5}$$

Define the marginal values of units in terms of the true value $v = \max_P \min_\Pi H(P, \Pi)$, by $s_i = v - w(d_i)$, $\sigma_j = w(\delta_j) - v$. Note that we do not yet know whether $\min_\Pi \max_P H(P, \Pi) = v$ or not. But from (5) and definitions we get

$$\max_P \min_\Pi \left(\sum_j f_j(P)\,\sigma_j - \sum_i \phi_i(\Pi)\, s_i\right) = 0. \tag{6}$$

Writing H as $\frac{vF(P) + a(P) + v\Phi(\Pi) - b(\Pi)}{F(P) + \Phi(\Pi)}$ we see that $\max_P a(P) = \max_\Pi b(\Pi)$ and that v is a true minimax value. Precisely, from (6), P is optimal if and only if it maximizes the linear function $\sum_j f_j(P)\,\sigma_j$ (and similarly for Π). Taking account of the constraints 1.2, we have

$$\sum_j \alpha_j \rho_j \sigma_j + \sum_i r_i \max_j b_{ij}\sigma_j = \sum_i a_i r_i s_i + \sum_j \rho_j \max_i \beta_{ji} s_i, \tag{6'}$$

and optimal fire distributions are pure distributions. Each friendly type i selects that enemy type j which maximizes $b_{ij}\sigma_j$, and directs all its fire upon that type.

Again the computation hinges on approximating the true value v, and consists of solving linear problems, one of which is "tangent" to the true problem at the solution. For any proposed value w, one solves (6') (quite rapidly; it consists of some arithmetic and choice of maxima from several rows of numbers) and commonly finds inequality instead of the equality written in (6'). If equality holds, then w is the true value. If the left side is greater, then w is too small. In this case the firing strategies which solve (6') for that w, when substituted into $H(P, \Pi)$, will yield a larger trial value, w'. If one proceeds with w', then the iteration will necessarily terminate; but we have not sufficient numerical experience or understanding of the computation to claim that this is the best method. After all, now that we have

proved that optimal pure distributions of fire exist, the examination of all pairs of pure strategies is a finite method. That proposed above, however, has the advantage of stepping monotonically to the solution.

Let us note that one could deduce the minimax theorem for $H(P, \Pi)$ from a more general minimax theorem for bilinear fractional functions [14]. The form $\frac{a + c}{b + d}$ permits us to solve by solving a succession of linear programming problems. It is the special form of the constraints (1.2) which trivializes these linear problems. Now, among the numerous respects in which our model fails to describe real combat accurately, one may note that we have admitted the possibility of forty battleships converging upon a rowboat to sink it in a fraction of a second. This can be corrected to some extent by imposing additional constraints. The cost is the introduction of linear programming problems more difficult than choosing the maximum of a row of numbers.

3. EXTENSIVE PROGRAMMING GAMES

One can combine fire programming problems with Hotspot assignment problems as treated in Part II. The assumptions above are similar to those in the third model of Part II. If we try to build on either of the first two models, we find new difficulties. One must decide when changes in fire distribution can be made. The main alternatives are (1) whenever a unit is lost, and (2) at the same times as the assignment changes. If we follow (1), then we must make a more careful analysis of the attrition process than we have yet done (and as noted in Part I, we are not able to handle the problem); if (2), then we face a more difficult type of fire programming game than that considered above. In either case, the computational problems appear to be beyond the present capacity of computers.

Accordingly, we return to the move rule of Section II.3: new assignments can be made precisely when a unit is lost. We shall not describe the model in detail; the commanders determine assignments, meet the enemy, and face a fire programming game involving the forces which have been assigned. The existence of a solution and the outlines of a computational method are fairly obvious; one must begin with forces of one unit and compute backwards up the compound sequence of minimax problems. The construction of such composite games, and the existence of solutions, are worked out in a very general setting in [6].

Even with the model we have just selected as the easiest to handle, there are serious difficulties. It does not seem reasonable to use the rules we have imposed on strategic decisions—that they change only when a unit is killed—unless one is lucky enough to have a model in which the commanders could not gain by disregarding these rules. On the other hand, these rules may be too permissive if they lead to a battle in which hundreds of units are engaged and strategies are changing several times a minute. We have not been able to investigate whether the second objection applies to the present models. But a certain battle of four units shows that the first objection does apply.

We define the example and give the pure strategies involved in best play; the interested reader can reproduce the computations. The 2 x 2 firepower matrices are

$$(b_{ij}) = \begin{pmatrix} 5 & 1 \\ 1 & 2 \end{pmatrix}; \qquad (\beta_{ji}) = \begin{pmatrix} 1 & 15 \\ 1 & 1 \end{pmatrix}$$

All a_i and α_i vanish. The final value is 1 if the enemy is destroyed (or quits the fight), 0 otherwise. The game values $w(m_1, m_2; \mu_1, \mu_2)$, when m_i units of the i-th friendly type (and μ_j for the enemy) are available, are these:

$$w(1, 0;\ 1, 0) = .8333$$
$$w(1, 0;\ 0, 1) = .5000$$
$$w(0, 1;\ 1, 0) = .0625$$
$$w(0, 1;\ 0, 1) = .6667$$
$$w(1, 0;\ 1, 1) = .3571$$
$$w(0, 1;\ 1, 1) = .0392$$
$$w(1, 1;\ 1, 0) = .8661$$
$$w(1, 1;\ 0, 1) = .8750$$

All available units are assigned to the battle in these eight cases. In the first four there is no choice of targets; in the next two the first enemy unit is the target; in the last two the first and second friendly unit, respectively, are the preferred targets. Note that in (1, 1; 1, 0) the marginal value of the first friendly unit is more than fifteen times as great as the marginal value of the second friendly unit.

Coming to the case (1, 1; 1, 1), there are five non-trivial fire programming games to be solved. In each case, when a unit has a choice of targets, the choice is made according to the strategies

$$P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}; \qquad \Pi = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

The game matrix, labeled with assignment strategies, the rows for the friendly force, is

	(1, 1)	(0, 1)	(1, 0)
(1, 1)	.5001	.6594	.5051
(1, 0)	.6362	.4527	.7357
(0, 1)	.4137	.6964	.3895

There is no saddle point. The minimax in pure strategies is .6362, but the maximin is .5001. For mixed strategies, the unique kernel is the 2 x 2 submatrix in the upper left; the value is .5633.

This means that whatever assignment the commanders make, as soon as each identifies the opposing force, one commander will see that he could gain either by sending a unit into reserve or by taking one from reserve. Under these conditions stability will exist, in the physical situation; but to find it we must specify the conditions of the battle much more fully, and the solution of the problem will be correspondingly more difficult.

PART IV. VALUE ESTIMATES

In this part we propose estimates of intrinsic values of weapons, guided by the models already examined. For the case of one type of weapon against one enemy type, as described

above, we find two simple formulas which seem to be useful in the respective cases $b\beta \geqq a\alpha$ and $b\beta \leqq a\alpha$; the formulas coincide when $b\beta = a\alpha$. Each formula is most useful when the friendly and enemy forces are about evenly matched. Thus the general problem of estimation of values is quite complex, even when only single types are considered. When several types are to be compared, one can begin by comparing them one against one. We find values in the mixed problem which seem to be a reasonable generalization of the "main" case $b\beta > a\alpha$, again provided the forces are about evenly matched. However, we find them only in that we prove existence and (except in certain boundary cases) uniqueness; an effective computational method remains to be found.

1. THE HOMOGENEOUS CASE

Recall the formula

$$\rho = \frac{\alpha - a + \sqrt{(\alpha - a)^2 + 4b\beta}}{2\beta} r . \tag{I.3.2}$$

This is the equation of the one straight line which is a flow line for the Lanchester equations. It suggests the attribution of values to friendly and enemy units in the ratio

$$\alpha - a + \sqrt{(\alpha - a)^2 + 4b\beta} \,/ 2\beta .$$

Let v and w be values in this ratio. Then we can say, in a Lanchester attrition process, that if $rv > \rho w$ then the friendly force wins, in the reverse case the enemy force wins, and in the case of equality the battle is a draw. The margin of victory depends non-linearly on the forces, according to a (generalized) Law of n^2; and thus the assignment of linear homogeneous values to units cannot be a completely satisfactory description. However, one is interested in estimating values according to simple formulas, for obvious reasons of economy; and the above formula seems convincing for the Lanchester model.

The formula (I.3.2) arises from solving the equation $\dot{r}/\dot{\rho} = r/\rho$ — a "stability" condition. Let us notice that the same ratio arises from the equations

$$\begin{aligned} \lambda v &= bw - av \ ; \\ \lambda w &= \beta v - \alpha w . \end{aligned} \tag{1}$$

In (1) we demand that the value v of a friendly unit be proportional to the net damage this unit inflicts upon the enemy; in time dt it destroys enemy units worth bw dt, but also destroys parts of itself worth av dt. It turns out also that $\lambda > 0$ if, and only if, $b\beta > a\alpha$. In fact, if $b\beta < a\alpha$ then the marginal value of a unit added to the battle is negative (if we assume that this unit could be kept in reserve, but thrown into the battle when needed). One can easily work this out with Lanchester equations; but, as must be clear, we wish to propose the use of (1) and (I.3.2) in other models and particularly in the model of Section II.3.

If $b\beta \geqq a\alpha$, then we know optimal strategies in Section II.3; the strategies consist of putting all available units into the battle, and it is worth emphasizing that limitations on movement of units, which would complicate the other case, leave this case unaffected. Using optimal strategies, we have a completely defined stochastic process. This process has been investigated by Richard Brown [2]; and it follows easily from Brown's results that, using v and w as given by (1):

THEOREM 1 (Brown): Asymptotically as r and ρ increase, if $rv > (1+\epsilon)\rho w$ for some $\epsilon > 0$, then the friendly force wins with probability 1; in the opposite case, the enemy wins with probability 1.

Brown has among other results a closer study of asymptotic behavior near the critical ratio, but it does not seem that we have a good enough approximation to reality to justify much concern over this matter. More to the point is the question whether for small forces the critical ratio accurately describes the advantage; if $rv > \rho w$, is the probability of winning greater than the probability of losing? Not always; however, several hundred values have been computed, and the description appears to be rather highly accurate.

Brown's results may apply as well when $b\beta < a\alpha$; but our investigations have shown that if commanders have some freedom to move units to and from a safe reserve, then Brown's stochastic process should not occur. What should occur, in the ideal case of II.3, is a sequence of duels between single units of each force. Now David Blackwell has pointed out that this is a sequence of independent events with identical probability distributions; and from the strong law of large numbers [5, p. 208], one easily deduces the same theorem, except that the critical ratio is different. The ratio is still determined by $\dot{r}/\dot{\rho} = r/\rho$, and turns out to be $(\alpha + b)/(a + \beta)$. Summarizing,

Definition: If $b\beta \geq a\alpha$, the values v of one friendly unit and w of one enemy unit are in the ratio $v/w = \dfrac{\alpha - a + \sqrt{(\alpha - a)^2 + 4b\beta}}{2\beta}$; if $b\beta \leq a\alpha$, then $v/w = \alpha + b/a + \beta$.

THEOREM 2 (Blackwell): For a sequence of duels, asymptotically as r and ρ increase, if $rv > (1+\epsilon)\rho w$ for some $\epsilon > 0$, then the friendly force wins with probability 1; in the opposite case, the enemy wins with probability 1.

We can deduce from the theorems of Brown and Blackwell and the fact, proved in Section II.3, that when $b\beta = a\alpha$ all non-zero assignments are equivalent, the two formulas must coincide in this case. In fact one easily sees that both reduce to $v/w = b/a$.

For small forces in Blackwell's case, we know of no extensive computations; it appears that the formula ought to determine the probable winner, at least as accurately as Brown's formula in Brown's case.

The formulas in both cases are found by solving $\dot{r}/\dot{\rho} = r/\rho$. We cannot generalize this to m types of weapons vs. n types (to obtain values; actually one can obtain critical lines, but we want a critical hypersurface). We can and shall generalize the equations (1), which also give the correct formula in Brown's case. The arguments for the generalization are somewhat tenuous, and we cannot even describe its range of approximate validity. We have no asymptotic theorems for non-homogeneous forces, nor do we have a full description of flow lines for generalized Lanchester equations. Thus the main support for the generalization is the superficial plausibility of (1).

Let us notice three criticisms applicable to our arguments above. Each, with our answer, can easily be reformulated for non-homogeneous forces.

The value of a unit in a battle is conceptually divisible into two parts; there is the immediate value of its firepower, applied to the destruction of enemies, and there is a survival value, equally relevant in our models whether the unit is in battle or in reserve. When $a\alpha > b\beta$ there is a facile distinction between these values; the "immediate" value is negative, and extra units should be held in reserve. The equations (1) seem to measure only this part of the value. On the other hand, when $a\alpha > b\beta$ and we permit the battle to be fought as a sequence of duels,

it seems that the "immediate" value vanishes and the formula $\frac{\alpha + b}{a + \beta}$ measures the "survival" value correctly. This argument suggests that further analysis should reveal other considerations which bear on evaluation in the case $b\beta > a\alpha$ — the main case, and the very case which we wish to generalize. The authors do not know what to do about this.

Second, notice that, although we are working without reference to the precise final conditions, i.e., the values of the states in which the battle may end, there is a quite general (if vague) property which these final conditions must have. There must be a substantial value attached to winning the field; otherwise the inferior force would simply refuse battle. This means that marginal values of units are greatest when the forces are about evenly matched, since, for example, the accession of one friendly unit in this case means a substantial increase in the probability of ultimate victory. In contrast, adding one unit to an inferior force merely decreases the expected size of the victorious enemy force and increases the probability of victory, perhaps from .001 to .002. (We pause to note: (1) the remark "the inferior force would. . .refuse. . .battle" can be proved as a theorem under reasonable hypotheses; (2) the remark "the marginal values. . .are greatest when the forces are about evenly matched" can probably be proved, but here we assert it on the basis of numerical experience.) To continue, it follows that in the case of evenly matched forces, a large part of the value of units is dependent on final conditions, i.e., on a factor which we have ignored. But not only is this case a specially important one, it also is one which we have given prominence in our arguments (the straight flow line, the theorems of Brown and Blackwell).

Our answer to this criticism is, first, that it limits the usefulness of our formulas. The formulas cannot be expected to measure marginal values accurately, whether the forces are evenly matched or not. They may be expected to measure average values. Second, we have the well-known argument from ignorance: the objection does not suggest that our ratio is too high, nor that it is too low, and therefore if the final conditions are roughly symmetric then we may hopefully use the proposed formulas until someone proposes better formulas.

The reader may assess the weights of the objection and answer just above. Let us notice two possible sources of asymmetry in the final values, cost or difficulty of replacing lost weapons, and intrinsic value of human life, both of which may be measured differently for the two forces.

The third criticism is more detailed and less significant, but it may be illuminating. In (1), if $a = \alpha$ then the magnitude of their common value is irrelevant (λ depends on a, but v and w are unaffected because $\lambda + a$ remains constant). But if $b = 4$, $\beta = 1$, $a = \alpha = 0$, then one friendly unit has a four-to-one advantage over one enemy; if $a = \alpha = 2$, the advantage is reduced to two-to-one. Is this consistent? The answer is that our (linear) formulas do not measure the advantage, which is not even approximately linear. What the formulas aim to measure is the numerical ratio of forces required to overcome a disadvantage in firepower. We are once more in the position of approximating a curved line with a straight line which is tangent at the critical point.

Turning to the example, the ratio $\alpha - a + \sqrt{(\alpha - a)^2 + 4b\beta}$ to 2β is two-to-one; we thence conclude that two enemies are about equally matched with one friendly unit. In fact, in the first case ($b = 4$, $\beta = 1$, $a = \alpha = 0$, $r = 1$, $\rho = 2$) we have 8 to 7 odds for winning, and in the second case it is 11 to 10 for the enemy. If a and α are increased further, and remain equal, we shall have $a\alpha > b\beta$ and must use the other formula. If $a = \alpha = 5$, the formula says that two friendly units equal three enemies; computing the probability of winning in this case,

we find it to be .4752. This is a fair selection of the examples we have computed, i.e., the formulas seem to be about as accurate as these three examples suggest. The accuracy of the model as a representation of battle is a different problem.

2. THE VALUE EQUATIONS

We propose to estimate values of units by using the equations

$$(\lambda + a_i)\, v_i = \max_j b_{ij} w_j ;$$

(1)

$$(\lambda + \alpha_j)\, w_j = \max_i \beta_{ji} v_i .$$

We have described our procedure loosely as approximating a curved line with a straight line. So far it seems more like approximating a curved line with a pound of butter, for there is little formal analogy between a stochastic process and a system of equations such as (1). In fact we can exhibit a stochastic maximizing problem to which (1) is closely related.

First, let us notice an inferior alternative. We have already distinguished between the "immediate" value of a unit, contributed by use of its firepower in the period of time before the first casualty, and the residue, which we called the "survival" value. One can define the immediate value precisely, by comparison with the alternative process in which the one unit in question cannot be used until after the first casualty occurs. The immediate value may be negative, and in any case should be much smaller than the survival value, for large numbers of units. Yet, as we said above, our equations seem to get at something like the immediate value. Therefore, one may expect that a model resembling that of Part III, but for which (1) is a perfectly accurate determinant of values, will involve some scheme for magnification of the immediate value. At any rate, our model below does magnify; it does this by repeating the period before the first casualty for an indefinite time.

Let m weapons and n targets be given. Each weapon is to choose a target and fire at it until the weapon blows up or the referee stops the game. A hit on the j-th target scores w_j points but does not destroy the target. The referee lets play continue for time t or more with probability e^{-kt}. In an interval dt, if the i-th weapon is firing at the j-th target, it has probability $a_i\, dt + o(dt)$ of blowing up and $b_{ij}\, dt + o(dt)$ of scoring a hit. Then one easily verifies that the i-th weapon should choose its target j so as to maximize $b_{ij}\, w_j$; the expected total score is $\sum_i \frac{1}{k+a_i} \max_j b_{ij} w_j$. Thus the value of the game is the sum of the values v_i of the units, where $(k+a_i)\, v_i = \max_j b_{ij} w_j$.

We have now presented all the arguments we know in favor of the estimate we propose, with several indications of conditions limiting the usefulness of the estimate. Let us repeat that we expect it to be useful over some range of parameters generalizing the range $b\beta \geq a\alpha$. Perhaps it is "valid" when all units are committed to the battle; this is almost surely not the best strategy when $\lambda < 0$. (In the example in Section III.3, $\lambda = \sqrt{15} - 1 > 0$. The values from (1) are 6 to 5.16 in favor of the friendly force. Yet if all units went into battle the odds would be almost exactly even.) However, it is reasonable to suppose (and it is taught) that usually the best strategy is to commit all available force to the battle; and frequently it is not even possible to keep a safe reserve. Thus our proposed estimate ought to describe, with fair

accuracy in most cases, (a) whether or not the friendly force is superior to the enemy, and (b) the comparative values of different components, or projected components, of the friendly force.

There remains the problem of solving (1), which presents considerable difficulties. Combining the equations, we obtain

$$v_i = \frac{1}{\lambda + a_i} \max_j \frac{1}{\lambda + \alpha_j} b_{ij} \max_k \beta_{jk} v_k . \tag{2}$$

THEOREM 1: Let a_i, b_{ij}, α_j, β_{ji} be non-negative real numbers such that for each i some $b_{ij} > 0$ and for each j some $\beta_{ji} > 0$. Then there exist a non-negative non-zero vector v and a real constant λ such that (2) holds.

PROOF: Consider the unit simplex S consisting of all v such that $v_i \geq 0$ for each i and $\sum_i v_i = 1$. Let R be the set of all real numbers greater than the maximum of all $-a_i$ and $-\alpha_j$. For v in S and λ in R, define $T(v,\lambda) = v'$, where $v'_i = \frac{1}{\lambda + a_i} \max_j \frac{1}{\lambda + \alpha_j} b_{ij} \max_k \beta_{jk} v_k$. Let $f(v,\lambda) = \sum_i v'_i$. For fixed v, if $\lambda' > \lambda$, then $f(v,\lambda') < f(v,\lambda)$. We omit the verification, which is routine but involves a tangled string of symbols; the inequality holds for each coordinate. Now as $\lambda \to \infty$, $f(v,\lambda) \to 0$; as λ approaches its lower limit, $f(v,\lambda)$ increases to infinity. Therefore, there exists just one $\lambda = \lambda(v)$ such that $f(v,\lambda) = 1$, i.e., $T(v,\lambda)$ is in S. We omit the routine verification that $\lambda(v)$ depends continuously on v. Now let $\Phi(v) = T(v, \lambda(v))$; Φ is a continuous transformation of S into itself, and hence by the Brouwer fixed point theorem there is v^0 such that $\Phi(v^0) = v^0$. But this means that v^0 and $\lambda(v^0)$ solve (2).

COROLLARY: If all b_{ij} and β_{ji} are positive then all v_i are positive.

PROOF: Some v_i, say v_1, is positive. Then for each i,

$$v_i \geq \frac{1}{\lambda + a_i} \frac{1}{\lambda + \alpha_1} \cdot b_{i1} \beta_{11} v_1 > 0 .$$

The corollary is not true without the special hypothesis. Consider $a_i = \alpha_j = 0$ for $i = 1, 2$, $j = 1, 2$; $b_{11} = b_{22} = \beta_{11} = 1$, $\beta_{22} = 2$, b_{ij} and β_{ij} vanish for $i \neq j$. One may verify that the above corollary, as well as Theorem 2 and some further remarks below, fails for this example. In fact the example amounts to two unconnected battles which cannot have the same value of λ; and we encounter another limitation on the usefulness of our estimate. When optimal strategies break up the battle into separate engagements, it is probably appropriate to deal with the parts separately. How to combine the results will, of course, depend on what is to happen after the separate engagements; we have no suggestions to offer.

For the remainder of the paper we assume that all b_{ij} and all β_{ji} are positive. We shall also write the extra conditions which we have satisfied:

$$\sum_i v_i = 1 ;$$

$$v_i > 0 \text{ and } w_j > 0 \text{ for all } i, j ; \tag{3}$$

$$\lambda > -a_i \text{ and } \lambda > -\alpha_j \text{ for all } i, j .$$

Given v and λ, w is obviously unique and positive. The condition $\sum_i v_i = 1$ enables us to speak of uniqueness of v.

THEOREM 2: In the solution of (2) and (3), λ is unique.

PROOF: Suppose (v, λ) and (v', λ') are solutions. We may assume $\lambda' > \lambda$, $v'_1 > v_1 > 0$, and the ratio $t_i = v'_i/v_i$ is maximum at $t_1 = t$. We shall derive the contradiction that $v'_k \geq \theta\, t v_k$ for some k and a certain $\theta > 1$. Choose k so that for some j,

$$v'_1 = \frac{1}{\lambda' + a_1} \frac{b_{1j}\beta_{jk}}{\lambda' + \alpha_j} v'_k .$$

Define $\theta = \dfrac{\lambda' + \alpha_j}{\lambda + \alpha_j} \dfrac{\lambda' + a_1}{\lambda + a_1}$; since $\lambda' > \lambda$ and the numerators and denominators involved are positive, therefore $\theta > 1$. Now suppose that $v'_k < \theta\, t v_k$; we conclude

$$(\lambda' + a_1)\, v'_1 = \frac{b_{1j}\beta_{jk}}{\lambda' + \alpha_j} v'_k < \theta\, t \frac{b_{1j}\beta_{jk}}{\lambda' + \alpha_j} v_k$$

$$\leq \frac{\theta t}{\lambda' + \alpha_j} (\lambda + \alpha_j)(\lambda + a_1)\, v_1 = (\lambda' + a_1)\, t v_1 ,$$

contradicting the definition of t. Hence $v'_k > t v_k$, another contradiction, which proves the theorem.

Next we shall note that λ satisfies at least one algebraic equation in the a's, α's, b's and β's; and we shall show that if v is not unique then λ satisfies at least two such equations involving disjoint sets of parameters. It follows that v is non-unique only when the parameters satisfy certain algebraic equations over the rationals. But if the parameters be chosen at random, the probability that they are algebraically related is zero. Further, the set of parameter values for which v is non-unique is topologically non-dense; for only finitely many algebraic equations need be examined. Thus we shall conclude

THEOREM 3: The solution of the system (2), (3) is unique except for a non-dense set of parameter values which has measure zero.

We need to sort the indices, but it will suffice to examine the friendly indices, $i, k, \ldots$; let k be a consequent of i if for some j,

$$v_i = \frac{1}{\lambda + a_i} \frac{b_{ij}\beta_{jk}}{\lambda + \alpha_j} v_k .$$

Then in a finite set, each index has at least one consequent; it follows that there exists a cycle $i_1, \ldots, i_n$, such that i_{p+1} is a consequent of i_p for each p and i_1 is a consequent of i_n (possibly $n = 1$). By eliminating the v's around a cycle of length n, we find an algebraic equation

of degree $2n$ satisfied by λ. Now from the formulas used in proving Theorem 2 we can deduce the

LEMMA: If (v, λ) and (v', λ) solve (2), then for any positive real number t the set of all indices i such that $v_i' \geq t\, v_i$ contains all consequents of its elements relative to v'.

Dually, the set of all i such that $v_i > \frac{1}{t}\, v_i'$ contains all consequents of its elements relative to v. If v and v' are not proportional, then t can be chosen so that both of these sets are non-empty. Then we have two cycles, and two independent algebraic equations satisfied by λ. Furthermore, there are only finitely many equations to be examined (the highest degree is twice the number of indices). Therefore Theorem 3 is proved.

THEOREM 4: The solution of (2) is a lower semi-continuous function of the parameters.

This merely says that the equations are continuous, i.e., a limit of solutions is a solution of the limit, which is obvious. One easily deduces the

COROLLARY: The solution of (2) and (3) is almost everywhere unique and continuous.

Continuity falls only on the bounding surfaces on which the solution is not (usually) single-valued. In fact, we can describe these surfaces as the places where the battle decomposes into unconnected parts — an unpleasant possibility which we mentioned above. We omit the tedious formalization; the following statement is unambiguous and the interested reader can continue the argument of Theorem 3 so as to prove the

Assertion: The values are non-unique if and only if the battle decomposes into two parts so that no weapon in either part has an optimal target in the other part.

Let us emphasize that the domain in which the present values are inappropriate includes not just the singular surfaces but some unknown neighboring regions.

Having proved the existence and frequent uniqueness of the solution of (2) and (3), we naturally inquire how to compute the solution. The analytic argument of Theorem 1 can be followed by approximate computational methods, but the labor involved is beyond all reason. The best method we have is a well-known, general, and thoroughly unsatisfactory one: guess values of v_i and a value of λ, substitute them into the right side of (2), obtain new values on the left side, fudge up a new value of λ somehow, and continue until the error is small or the computer is exhausted. We have an example in which, given the correct value of λ but a bad guess at v, this method will not converge; however, the method usually works and is fairly rapid for small numbers of weapon types.

For the important special case in which all a_i and all α_j vanish, there is a finite but impractical method which is described by one of the authors in [7]. Progress in the study of these estimates will be facilitated by the development of better computing methods.

391368 O - 56 - 7

REFERENCES

[1] Bellman, R., "On an iterative procedure for obtaining the Perron root of a positive matrix," Proc. Amer. Math. Soc. Vol. 6 (1955), pp. 719-725.

[2] Brown, R. H., A stochastic analysis of Lanchester's theory of combat, Operations Research Office Technical Memorandum, ORO-T-323, 1955.

[3] Dines, L. L., "On a theorem of von Neumann," Proc. Nat. Acad. Sci. U.S.A., Vol. 33 (1947), pp. 329-331.

[4] Feller, W., "On the theory of stochastic processes," Proceedings of the Berkeley symposium on mathematical statistics and probability, Berkeley, 1949.

[5] _____, An introduction to probability theory and its applications, New York, 1950.

[6] Isbell, J. R., "Finitary games," to appear in an Annals of Mathematics Study, Contributions to the theory of Games, Vol. III.

[7] _____, "A stability problem in linear algebra," Ordnance Computer Research Report, Vol. 3 (1956), pp. 24-26.

[8] Loomis, L. H., "On a theorem of von Neumann," Proc. Nat. Acad. Sci., U.S.A., Vol. 32 (1946), pp. 213-215.

[9] Morse, P. M. and Kimball, G. E., Methods of operations research, New York, 1951.

[10] Shapley, L. H., "Stochastic games," Proc. Nat. Acad. Sci. U.S.A., Vol. 39 (1953), pp. 1095-1100.

[11] Snow, R. N., Contributions to Lanchester attrition theory, Project Rand Report No. RA-15078, 1948.

[12] Tompkins, C., "Probabilistic problems and military evaluation: an example," The George Washington University Logistics Papers, Issue 2, 1950.

[13] Truesdell, C. A., An essay toward a unified theory of special functions, Princeton, 1948.

[14] von Neumann, J., Ergebnisse eines Math. Kolloquiums, Vol. 8 (1937), pp. 73-83; in translation: Review Economic Studies, Vol. 13 (1945-46), pp. 1-9.

* * *

AN ALGORITHM FOR QUADRATIC PROGRAMMING

Marguerite Frank and Philip Wolfe[1]
Princeton University

A finite iteration method for calculating the solution of quadratic programming problems is described. Extensions to more general non-linear problems are suggested.

1. INTRODUCTION

The problem of maximizing a concave quadratic function whose variables are subject to linear inequality constraints has been the subject of several recent studies, from both the computational side and the theoretical (see Bibliography). Our aim here has been to develop a method for solving this non-linear programming problem which should be particularly well adapted to high-speed machine computation.

The quadratic programming problem as such, called PI, is set forth in Section 2.

We find in Section 3 that with the aid of generalized Lagrange multipliers the solutions of PI can be exhibited in a simple way as parts of the solutions of a new quadratic programming problem, called PII, which embraces the multipliers. The maximum sought in PII is known to be zero. A test for the existence of solutions to PI arises from the fact that the boundedness of its objective function is equivalent to the feasibility of the (linear) constraints of PII.

In Section 4 we apply to PII an iterative process in which the principal computation is the simplex method change-of-basis. One step of our "gradient and interpolation" method, given an initial feasible point, selects by the simplex routine a secondary basic feasible point whose projection along the gradient of the objective function at the initial point is sufficiently large. The point at which the objective is maximized for the segment joining the initial and secondary points is then chosen as the initial point for the next step.

The values of the objective function on the initial points thus obtained converge to zero; but a remarkable feature of the quadratic problem is that in some step a secondary point which is a solution of the problem will be found, insuring the termination of the process.

A simplex technique machine program requires little alteration for the employment of this method. Limited experience suggests that solving a quadratic program in n variables and m constraints will take about as long as solving a linear program having $m+n$ constraints and a "reasonable" number of variables.

Section 5 discusses, for completeness, some other computational proposals making use of generalized Lagrange multipliers.

Section 6 carries over the applicable part of the method, the gradient-and-interpolation routine, to the maximization of an arbitrary concave function under linear constraints (with one qualification). Convergence to the maximum is obtained as above, but termination of the process in an exact solution is not, although an estimate of error is readily found.

In Section 7 (the Appendix) are accumulated some facts about linear programs and concave functions which are used throughout the paper.

---

[1]Under contract with the Office of Naval Research.

2. THE QUADRATIC PROBLEM

Our primary concern is with the problem of maximizing the concave quadratic function

$$f(x) = \sum_{j=1}^{n} p_j x_j - \sum_{j,k=1}^{n,n} x_j C_{jk} x_k \tag{2.1}$$

subject to the linear constraints

$$\begin{aligned} x_j &\geq 0 \qquad (j = 1, \ldots, n), \\ \sum_{j=1}^{n} A_{ij} x_j &\leq b_i \qquad (i = 1, \ldots, m). \end{aligned} \tag{2.2}$$

Matrix notation will be used exclusively below. x is the $n \times 1$ matrix (i.e., column vector) of variables $x_1, \ldots, x_n$. A, C, p, and b are, respectively, $m \times n$, $n \times n$, $1 \times n$, and $m \times 1$ matrices. $A_{i.} = (A_{i1}, \ldots, A_{in})$ will denote the i^{th} row of A, and likewise $A_{.j}$ the j^{th} column. e_j is the j^{th} coordinate (column) vector of n dimensions. The symbol $'$ denotes matrix transposition. For any function $f(x)$, the gradient $1 \times n$ matrix $\left(\frac{\partial f}{\partial x_1}, \ldots\right)$ is denoted by ∂f. C may without loss of generality be supposed symmetric.

In matrix terms, we may rewrite the quadratic problem as

PI: Maximize $f(x) = px' - x'Cx$ subject to

$$(I) \quad \begin{cases} x \geq 0 \\ Ax \leq b. \end{cases}$$

An x satisfying (I) will be called feasible. The set of all feasible x is the constraint set. The problem PI is called feasible if the constraint set is not empty, that is, if there exists a feasible x. If the objective function has a supremum on the constraint set, this supremum will be denoted by M. An x for which $f(x) = M$ is called a solution of PI and the set of all solutions will be called the solution set. It will be assumed in this and the following section that PI is feasible, but not that solutions necessarily exist.

The postulated concavity of the objective function is equivalent to the requirement that C be the matrix of a positive semidefinite quadratic form [App. d]. A function is said to be concave if interpolation never overestimates its values: that is, if

$$f(\alpha x + \beta y) \geq \alpha f(x) + \beta f(y) \text{ whenever } \alpha, \beta \geq 0,\ \alpha + \beta = 1. \tag{2.3}$$

It follows that any local maximum point x of f on a convex constraint set is also a global maximum: for otherwise a y such that $f(y) > f(x)$ could be found, and then any point on the segment joining x to y, no matter how close to x, would yield a higher value for f than would x. Similarly for a convex function, any local minimum on a convex set is a global minimum. It is this fact which makes maximizing a concave function (or minimizing a convex function) one of the simpler of the non-linear programming problems.

Concavity of the objective function implies that a "mixture" of two equally good sets of activities can never result in a poorer set of activities. Conversely, the worst possible program

is always "extreme" in that, roughly speaking, it involves as little mixing as is possible. In certain cases it may happen that an "extreme" program is also the best program; but this may be considered as not a typical case, although not a pathological case either. It appears that non-linear programming problems in which (2.3) does not hold — e.g., contract award programs if discounts are given for large orders — will require substantially different techniques than those presented here.

3. LAGRANGE MULTIPLIERS

Kuhn and Tucker's generalization [6] of the method of Lagrange multipliers to the maximization of a function f whose variables $x_1, \ldots, x_n$ lie in the constraint set given by the inequalities

$$f_i(x) \geq 0 \qquad (i = 1, \ldots, m) \tag{3.1}$$

is based on the observation that f has a local extreme value at x if, and only if, the normal hyperplane to the gradient vector $\partial f(x)$ at x is, locally, a supporting hyperplane of the constraint set. (This formulation applies equally well to the classical case $f_i(x) \equiv 0$.) If x is to be a local maximum, then the constraint set and the gradient drawn from x must lie on opposite sides of the supporting hyperplane. If, furthermore, the constraint set is convex (as when the functions f_i are convex or, in particular, linear), and f is concave, then, as seen in Section 2, this condition is necessary and sufficient that f have a global maximum at x.

Noting that a hyperplane touching the convex constraint set at a point x satisfies the above condition if, and only if, x has the maximum projection (of all points of the convex) along the outward normal to the hyperplane, a necessary and sufficient condition that x_0 be a solution of PI is that

$$\partial f(x_0)x_0 = \text{Max}\,[\partial f(x_0)w \mid w \geq 0,\ Aw \leq b], \quad (w \text{ an } n\times 1 \text{ matrix}). \tag{3.2}$$

But by the Duality Theorem for linear programming [App. a] (x_0 is fixed), the right side of (3.2) is equal to

$$\text{Min}\,[ub \mid u \geq 0,\ uA \geq \partial f(x_0)] \quad (u \text{ a } 1\times m \text{ matrix}). \tag{3.3}$$

Substituting this for the right side of (3.2) and transposing, we obtain that:

x_0 is a solution of PI if, and only if,

$$\text{Max}\,[\partial f(x_0)x_0 - ub \mid u \geq 0,\ uA \geq \partial f(x_0)] = 0\,. \tag{3.4}$$

Note that $\partial f(x) = p - 2x'C$. The function

$$g(x, u) \equiv \partial f(x)\,x - ub = px - ub - 2x'Cx \tag{3.5}$$

extracted from (3.4), which is similar to the generalized Lagrangian of [6], plays a dominant role in the sequel. It is evidently a concave function of the variables (x, u). We suppose henceforth that (x, u) is constrained by the linear relations used in PI, (3.2), and (3.4), namely,

$$\text{(3.6)} \qquad \begin{cases} x \geq 0, & u \geq 0, \\ Ax \leq b, & \partial f(x) \leq uA. \end{cases}$$

Now under these constraints

$$\text{(3.7)} \qquad g(x,u) \leq 0,$$

since $g(x,u) = \partial f(x)\, x - ub \leq uAx - ub = u(Ax - b) \leq 0$. But then the Max of (3.4) is never positive, and we have proved the result on which our use of the generalized Lagrange multiplier method rests, the

SOLUTION CRITERION:[2] x is a solution of PI if, and only if, for some u such that (x,u) satisfies (3.6),

$$\text{(3.8)} \qquad g(x,u) = 0.$$

More information connecting f and g is given by the following easy consequence of the concavity of f [App. e]:

For any x and y,

$$\text{(3.9)} \qquad f(y) - f(x) \leq \partial f(x)\,(y - x).$$

If now (x,u) satisfies (3.6), we have that for any feasible y,

$$\text{(3.10)} \qquad \begin{aligned} f(y) - f(x) &\leq \partial f(x)\, y - \partial f(x)\, x \leq uAy - \partial f(x)\, x \\ &\leq ub - \partial f(x)\, x \qquad = -g(x,u). \end{aligned}$$

Evidently if f is not bounded above for all feasible y, then such (x,u) cannot be found. On the other hand, [App. i] if f is so bounded then there exists a solution, x_0, and thus, by the Solution Criterion, (x_0,u) satisfying (3.6) exists. We have thus proved the

BOUNDEDNESS CRITERION: f is bounded above on the constraint set if, and only if, the joint constraints (3.6) are feasible.

Finally, supposing in (3.10) that y is a solution of PI—that is, that $f(y) = M$, we have the

APPROXIMATION CRITERION: If (x,u) satisfies (3.6), then

$$\text{(3.11)} \qquad M - f(x) \leq -g(x,u).$$

THE PROBLEM PII: From the above it appears that a solution x of PI, if there is any, may be read off from any (x,u) for which the maximum, zero, of the concave quadratic

---

[2] The Solution Criterion was originally obtained from Lemma 1 of [6]: our $g(x,u)$ is $\phi_x^{o\prime}\, x^o + \phi_u^{o\prime}\, u^o$ in Kuhn and Tucker's notation, so that (3.8) is equivalent to part of the conditions of their Lemma 1. (3.6) is equivalent to the remaining conditions, and Theorems 1 and 2 of [6] then establish the Solution Criterion. Barankin and Dorfman [1] obtain the Criterion and the transformed problem PII in this way.

function $g(x,u)$ is achieved under the constraints (3.6). It is, in fact, this latter problem, PII below, which is solved in Section 4. Also, the Approximation Criterion (3.11) yields an estimate $f(x) \leq M \leq f(x) - g(x,u)$ for M when any (x,u) satisfying (3.11) is at hand.

As is customary in solving linear programs [3, p. 339] the constraints (3.6) will be written as equations by adjoining one non-negative variable y_i $(i = 1,\ldots, m)$ to each of the m inequalities of $Ax \leq b$, and one v_j $(j = 1,\ldots, n)$ to each of the n inequalities of $\partial f(x) \leq uA$ (written for our purposes as $-\partial f(x) + uA \geq 0$, or, using matrix transposition, as $2Cx + A'u' \geq p'$), obtaining

$$
\begin{aligned}
&x \geq 0,\ u \geq 0,\ y \geq 0,\ v \geq 0\,, \\
&Ax + y = b\,, \\
&2Cx + A'u' - v' = p'
\end{aligned}
\tag{3.12}
$$

(y an $m \times 1$ matrix, v an $n \times 1$ matrix).

We also have, after substitution from (3.12),

$$
g(x,u) = \partial f(x)\, x - ub = (uA - v)\, x - u(Ax + y) = -(vx + uy)\,. \tag{3.13}
$$

In these terms, PII consists in finding vectors x, u, y, v satisfying (3.12) such that

$$
vx + uy = 0\,. \tag{3.14}
$$

This last condition may be interpreted geometrically by noting that it is equivalent to

$$
v_j \neq 0 \text{ implies } x_j = 0,\ u_i \neq 0 \text{ implies } y_i = 0 \text{ (all } i, j)\,. \tag{3.15}
$$

Since $\partial f(x) = uA - v = \Sigma\, v_j\,(-e_j) + \Sigma\, u_i\, A_{i.}$ is an expression for $f(x)$ as a non-negative linear combination of the outward normals to those bounding hyperplanes $e_j x = 0$, $A_{i.}\, x = b_i$ on which x lies (i.e., for which x_j, y_i vanish), the solution criterion (3.14) is equivalent to:

(3.16) x solves PI if, and only if, $\partial f(x)$ belongs to the convex cone spanned by the outward normals to the bounding hyperplanes on which x lies.

This condition is a geometric dual of the condition stated at the beginning of this section, and has been obtained from it by the duality theorem for linear programming.

BASIC SOLUTIONS: The system (3.12) consists of $m+n$ equations in $2(m+n)$ non-negative variables. It is easily shown that any extreme point of the set of points x, u, y, v feasible for the constraints (3.12) has at most $m+n$ positive components. An extreme point of this convex will be called a basic feasible vector.[3] Evidently the solution of a linear program can always be found as a basic feasible vector and the simplex process does precisely so [App. b].

[3] After Dantzig. Under the non-degeneracy hypothesis of [3], a basic feasible vector has exactly $m+n$ positive components.

Now the conditions (3.15) entail that at most $m+n$ components of any solution of PII are positive, so that we may hope to find a solution of PII among the basic feasible vectors of the constraint set for PII. We show in Section 4 that this is indeed the case.[4]

This remarkable consequence of the use of Lagrange multipliers might suggest the direct use of the simplex method to maximize the objective function; however, as discussed in Section 5, this is not possible. Nevertheless, the simplex method machinery is usefully exploited in the "gradient-and-interpolation" method presented in the next section.

4. THE COMPUTATION

PRELIMINARIES: For present purposes the four interdependent vectors x, u, y, v may be combined into a single constrained vector z. Accordingly let

$$(4.1) \qquad z = \begin{bmatrix} x \\ u' \\ y \\ v' \end{bmatrix}, \quad B = \begin{bmatrix} A & O & I & O \\ 2C & A' & O & -I \end{bmatrix}, \quad d = \begin{bmatrix} b \\ p' \end{bmatrix},$$

the two instances of I being identity matrices of appropriate rank. z, B, and d are then $2(m+n)\times 1$, $(m+n)\times 2(m+n)$, and $(m+n)\times 1$ matrices, respectively; and the system of constraints (3.12) for PII assumes the form.

$$(II) \qquad z \geq O, \quad Bz = d .$$

With each z as above associate the "adjoint"

$$(4.2) \qquad \tilde{z} = [v, y', u, x'] .$$

The new objective function may be conveniently expressed using this linear operation, for by (3.13)

$$(4.3) \qquad g(x,u) = -(vx + uy) = -\frac{1}{2}\tilde{z}\, z .$$

Further, for $z = [x', u, y', v]'$ and $Z = [X', U, Y', V]'$ feasible with respect to (II) above,

$$(4.4) \qquad \begin{cases} \tilde{z}\, Z = \tilde{Z}\, z , \text{ and} \\ (\tilde{Z} - \tilde{z})(Z - z) \geq 0 . \end{cases}$$

The first relation is evident. The second relation is given by

$$\begin{aligned} \frac{1}{2}(\tilde{Z} - \tilde{z})(Z - z) &= (V - v)(X - x) + (U - u)(Y - y) \\ &= [2(X - x)'C + (U - u)A](X - x) + (U - u)A(x - X) \\ &= 2(X - x)'C(X - x) \geq O , \end{aligned}$$

which thus reflects the concavity of g.

[4] This has been shown independently by Barankin and Dorfman when the system (3.12) is non-degenerate [1].

In these terms, the final version of the Lagrange multiplier problem obtained in the last section may be stated as

PII: Find z for which the maximum, zero, of the inner product

$$-\tilde{z}\,z$$

is assumed under the constraints

(II) $$z \geq 0\,, \quad Bz = d\,.$$

By the Boundedness Criterion, PI has a (finite) solution if, and only if, PII is feasible; by the Solution Criterion, x solves PI if, and only if, $z = [x'\ u\ y'\ v]'$ solves PII for some y, u, and v.

Although this has not been necessary before, for application of the simplex method we suppose that the constraint equations (II) have, where necessary, been multiplied by -1, so that the right side is non-negative.

THE ALGORITHM: The process below will yield a basic feasible vector solution of PII. Phase I initiates the process, while Phase II is iterated until it yields the required vector.

Phase I: The constraints (II) are tested for feasibility. (Most commonly employed is the artificial basis [App. c].) If they are feasible, a basic feasible vector z_1 is produced with which to begin Phase II. If not, the last n equations may be discarded and the remainder similarly tested for feasibility. If these are feasible, then the quadratic problem is feasible, but unbounded above; otherwise, the quadratic problem is infeasible.

Phase II: This phase is defined inductively: At the k^{th} instance of Phase II we have the feasible (II) vector w_k which does not solve PII, and also a basic feasible vector z_k with which to start the simplex machinery. (At the first instance, let $w_1 = z_1$.)

Employ the simplex technique in the maximization of the linear form

$$-\tilde{w}_k\,z\,,$$

obtaining the sequence of basic feasible vectors $z^1 = z_k,\ z^2,\ z^3, \ldots$ such that

$$-w_k z^1 < -w_k z^2 < \ldots\ ^5.$$

Stop at the first z^h such that either

(4.5) $$\tilde{z}^h z^h = 0\,, \quad \text{or}$$

(4.6) $$-\tilde{w}_k z^h \geq -\frac{1}{2}\tilde{w}_k w_k\,.$$

---

[5] In the likely event that the constraints (II) are degenerate, "$\leq$" may occur here for a while, but not for long. Dantzig's method for handling degeneracy [2, 4] is exceptionally easy to use here, owing to the presence (except for sign) of an identity matrix in the constraint matrix.

If (4.5) obtains, then z^h solves the problem. Otherwise let

$$(4.7)\qquad \begin{cases} z_{k+1} = z^h \\ \mu = \operatorname{Min}\left\{ \dfrac{\bar{w}_k (w_k - z_{k+1})}{(\bar{z}_{k+1} - \bar{w}_k)(z_{k+1} - w_k)} , 1 \right\}, \text{ and} \\ w_{k+1} = w_k + \mu (z_{k+1} - w_k) . \end{cases}$$

Repeat Phase II, using w_{k+1} and z_{k+1}.

PROOF OF CONVERGENCE: It will be shown that the new objective $-\bar{w}_k w_k$ converges monotonically to zero, and that in some instance of Phase II a basic feasible vector solution satisfying (4.5) is found.

(a) In order to show that z^h satisfying (4.6) will be found in a given instance of Phase II (if (4.5) is not obtained), it is sufficient to observe that the linear programming problem Max $\{-\bar{w}_k z \mid (II)\}$ has a finite maximum, since $-\bar{w}_k z \le 0$. This maximum is not less than $-\frac{1}{2}\bar{w}_k w_k$, since if Z solves PII, i.e., $\bar{Z}Z = 0$, we have $\bar{w}_k w_k - 2\bar{w}_k Z = (\bar{w}_k - \bar{Z})(w_k - Z) \ge 0$, by (4.4).

For the time being, let $w_k = w$, $z_{k+1} = Z$, $w_{k+1} = W$. Then (4.6) yields $\bar{w}(w - Z) \ge \frac{1}{2}\bar{w} w > 0$, and (4.4) that $(\bar{Z} - \bar{w})(Z - w) \ge 0$, so that $0 < \mu \le 1$ (no difficulty arises if the denominator in the expression for μ vanishes). Thus w_{k+1} is a convex combination of w_k and z_{k+1}; indeed, it has been chosen so as to maximize the new objective on the segment $\overline{w_k z_{k+1}}$.

(b) Convergence. Dropping subscripts as in (a), when $Z = z^h$ satisfying (4.6) is obtained in the k^{th} instance of Phase II, we have

$$\begin{aligned} \bar{W}W &= \bar{w}w + 2\mu\bar{w}(Z - w) + \mu^2(\bar{Z} - \bar{w})(Z - w) \\ &= \bar{w}w + \mu\bar{w}(Z - w) + \mu[\mu(\bar{Z} - \bar{w})(Z - w) - \bar{w}(w - Z)] \\ &\le \bar{w}w + \mu\bar{w}(Z - w) \qquad (4.8) \\ &\le \bar{w}w - \mu \cdot \frac{1}{2}\bar{w}w \\ &= (1 - \tfrac{1}{2}\mu)\bar{w}w . \end{aligned}$$

Letting F be the (compact) convex hull of the set of all basic feasible vectors of (II), let

$$(4.9)\qquad L = \operatorname{Max}\left\{ (\bar{z}^1 - \bar{z}^2)(z^1 - z^2) \mid z^1, z^2 \text{ in } F \right\}.$$

If $\mu < 1$, then

$$\frac{1}{2}\mu \geq \frac{\tilde{w}(w - Z)}{2L} \geq \frac{\tilde{w}w}{4L} \quad \text{by (4.6)},$$

so that, in any case,

$$1 - \frac{\mu}{2} \leq \text{Max}\left\{1 - \frac{\tilde{w}w}{4L}, \frac{1}{2}\right\},$$

and finally by (4.8)

$$\frac{\tilde{W}W}{4L} \leq \frac{\tilde{w}w}{4L} \text{ Max}\left\{1 - \frac{\tilde{w}w}{4L}, \frac{1}{2}\right\}. \tag{4.10}$$

Resuming subscripts, let $a_k = \frac{\tilde{w}_k w_k}{4L}$. Then (4.10) becomes

$$a_{k+1} \leq \text{Max}\left\{1 - a_k, \frac{1}{2}\right\} a_k \quad \text{for } k \geq 0. \tag{4.11}$$

Since if $a_k \geq \frac{1}{2}$ then $a_{k+1} \leq \frac{1}{2} a_k$, there exists K such that $a_k < \frac{1}{2}$ whenever $k \geq K$. Now for $a_k < \frac{1}{2}$ we have $a_{k+1} \leq a_k(1 - a_k)$, whence

$$\frac{1}{a_{k+1}} \geq \frac{1}{a_k} \frac{1}{1 - a_k} = \frac{1 + a_k + a_k^2 + \cdots}{a_k} \geq \frac{1}{a_k} + 1,$$

so that

$$\text{for } k \geq K, \quad \frac{1}{a_k} \geq \frac{1}{a_K} + (k - K).$$

Thus

$$a_k \leq \begin{cases} \left(\frac{1}{2}\right)^k a_0 & \text{for } k \leq K \\ \dfrac{1}{k - K + \frac{1}{a_K}} & \text{for } k > K. \end{cases} \tag{4.12}$$

Hence $\tilde{w}_k w_k$ approaches zero at worst like $1/k$, and the convergence is proved.

(c) Finiteness. Now suppose that the linear program of Phase II never yields a basic feasible vector which solves PII; then in particular no z_k obtained in the iteration of Phase II solves PII. Since each w_k is a convex combination of $\{z_{k'} \mid k' < k\}$, there is a point of

accumulation w_o of $\{w_k\}$, belonging to the convex hull of the finite set of basic feasible vectors $\{z_k \mid \bar{z}_k z_k > 0\}$ for which $\bar{w}_o w_o = 0$; yet $w_o = \Sigma t_k z_k$ with $t_k \geq 0$, $\Sigma t_k = 1$, so that $\bar{w}_o w_o = \sum_{j,k} t_j t_k \bar{z}_j z_k > 0$: a contradiction. The proof that Phase II terminates in a solution is complete.

(d) Approximation. Some iterations of Phase II can be avoided if only a solution of PI to a predetermined degree of approximation is desired; for the Approximation Criterion (3.11) yields that, where x_k is the x-part of w_k,

$$M - f(x_k) \leq \frac{1}{2} \bar{w}_k w_k ; \tag{4.13}$$

Phase II is repeated until $\bar{w}_k w_k$ is sufficiently small.

Computation Time: Experience with a few very small numerical examples (of the order of $n = 5$, $m = 3$) suggests that the total number of changes of basis (several in each instance of the application of the Simplex Technique to Phase II) required to solve the quadratic problem will, in practice, be of the order of $2(m + n)$. The instances of Phase II after the first commonly require only one or two changes of basis. Thus this algorithm may be able to solve quadratic problems as quickly as linear problems of comparable size can be solved; but we have no theoretical estimate for this.

5. OBITER DICTA

We briefly consider here some alternative computational proposals which aim directly to exploit the fact that some basic feasible vector of (II) is a solution of the problem PII by using the simplex change of basis to pass from a basic feasible vector to adjacent basic feasible vectors.

Perhaps the most attractive possibility is that of the existence of a function defined for basic feasible vectors which assumes its minimum at a solution and which can always be decreased with a single change of basis if the solution has not been reached. In view of the Solution Criterion (3.8), the function $-g(x, u) = \frac{1}{2} \bar{z} z$ naturally merits consideration from this angle, as well as the integer-valued function which counts the number of positive components of $\bar{z}$ that coincide in position with the positive components of z; these functions have been studied by Barankin and Dorfman [1]. Quadratic programming problems may be constructed, however, in which certain non-solving basic feasible vectors occur as "local minima" for these functions, which cannot be decreased with a single change of basis. Whether or not a function of the desired type exists is not known.

One could also systematically explore all the basic feasible vectors of (II) for a solution, changing basis to pass from vertex to adjacent vertex in the graph consisting of all the basic feasible vectors and their connecting edges in (II). Such a process has been proposed by Charnes [2] for obtaining all solutions of a linear programming problem, given one. The amount of information that must be recorded if a graph of unknown design is to be traced out has been studied as the "labyrinth problem" [5]. The tracing process becomes less efficient (in terms of edges retraced) the less the data recorded. A minimum record is a list of all vertices already traversed. The best available upper bound [7] for the number of vertices of (II) being of the order of $\binom{2m+2n}{m+n-2}$, it seems that such an approach to PII would be infeasible for large-scale problems because of the excessive demands made on the memory of a high-speed computer.

6. CONCAVE PROGRAMMING

The gradient-and-interpolation method of Section 4 applied to the quadratic concave function $g(x,u)$ may be equally well applied to the problem of maximizing a more general concave function f, satisfying the hypothesis (A) below, whose variables are constrained by linear relations. Generalized Lagrange multipliers do not seem to have the same utility here that they do in the quadratic case, since the constraint $\partial f(x) \leq u A$ entailed by their use is no longer linear, and thus the features which ensured the termination of the quadratic programming in a finite number of steps are not found here.

Let f be a concave function of n variables possessing continuous second derivatives. The concave programming problem is:

PC: Maximize $f(x)$ under the constraints

$$x \geq 0 , \quad Ax \leq b . \tag{6.1}$$

Here A, b are $m\times n$, $m\times 1$ matrices, $\partial f(x)$ is the $1\times n$ matrix $\left(\frac{\partial f(x)}{\partial x_i}\right)$, $\partial^2 f(x)$ is the $n\times n$ matrix $\left(\frac{\partial^2 f(x)}{\partial x_i \partial x_j}\right)$.

The problem will be supposed to satisfy the hypothesis

(A): For each feasible x, $\{\partial f(x)\, y \mid y \text{ feasible}\}$ is bounded above.

In particular, (A) holds if the constraint set is bounded, and in any case implies that f is bounded above on the constraint set, since for x fixed we have [App. e] the consequence of concavity

$$f(y) \leq f(x) + \partial f(x)\,(y - x) . \tag{6.2}$$

(A) is also easily shown for the problem PII.

Supposing now that the problem is feasible, let F be the (compact) convex hull of the set of basic feasible vectors of the constraint set. Such basic feasible vectors exist, since by (A) the linear program Max $\{\delta f(x)\, y \mid y \geq 0,\ Ay \leq b\}$ has solutions. There exists, then, a constant $L > 0$ such that for each x, y, and w in F,

$$-(y - x)'\, \partial^2 f(w)\, (y - x) \leq L . \tag{6.3}$$

The computational method follows.

Phase One: As in the quadratic case, determine the feasibility of the constraints (6.1), obtaining the initial basic feasible vector z_1.

Phase Two: At the k^{th} instance of Phase Two, when we suppose given the basic feasible vector z_k and the feasible vector x_k $(x_1 = z_1)$, find a basic feasible vector z_{k+1} which solves the linear program

$$\text{Max } \{\partial f(x_k)\, w \mid w \geq 0,\ Aw \leq b\} . \tag{6.4}$$

Let s be such that

$$-(z_{k+1} - x_k)' \, \partial^2 f(w) \, (z_{k+1} - x_k) \leq s \geq 0 \tag{6.5}$$

for all w on the segment $\overline{x_k z_{k+1}}$.[6] Let

$$\begin{cases} \mu = \text{Min} \left\{ \dfrac{\partial f(x_k) \, (z_{k+1} - x_k)}{s} , \, 1 \right\}, \\ x_{k+1} = x_k + \mu \, (z_{k+1} - x_k) , \end{cases} \tag{6.6}$$

and repeat Phase Two with z_{k+1}, x_{k+1}.

PROOFS: We shall prove that $f(x_k)$ converges to M, the supremum of f on the constraint set. Let for the moment $x_k = x$, $x_{k+1} = X$, and $z_{k+1} - x_k = y$.

From (6.2), for any feasible w we have $f(w) \leq f(x) + \partial f(x) \, (w - x) \leq f(x) + \partial f(x) \, (z_{k+1} - x) = f(x) + \partial f(x) \, y$, since z_{k+1} maximizes $\partial f(x_k)$; and thus

$$M - f(x) \leq \partial f(x) \, y . \tag{6.7}$$

Now by Taylor's theorem, since $0 \leq \mu \leq 1$,

$$f(X) = f(x + \mu y) = f(x) + \mu \, \partial f(x) \, y + \frac{1}{2} \mu^2 \, y' \, \partial^2 f(w) \, y ,$$

where w lies on $\overline{xX}$, so that, from (6.7), (6.6), and (6.5),

$$\begin{aligned} M - f(X) &= M - f(x) - \frac{1}{2} \mu \, \partial f(x) \, y - \frac{1}{2} \mu \, [\partial f(x) \, y + \mu \, y' \, \partial^2 f(w) \, y] \\ &\leq M - f(x) - \frac{1}{2} \mu \, [M - f(x)] - \frac{1}{2} [s \mu + \mu \, y' \, \partial^2 f(w) \, y] \\ &\leq (1 - \frac{1}{2} \mu) \, [M - f(x)] . \end{aligned} \tag{6.8}$$

Since, moreover,

$$1 - \frac{1}{2} \mu = \text{Max} \left\{ 1 - \frac{1}{2} , \frac{\partial f(x) \, y}{-y' \, \partial^2 f(w) \, y} \right\} \leq \text{Max} \left\{ \frac{1}{2} , \frac{M - f(x)}{L} \right\},$$

(6.8) yields, dividing by L, resuming subscripts,

---

[6] Of course $s \leq L$. For this method, s should be chosen as small as possible, but we have no suggestions for finding it.

$$\frac{M - f(x_{k+1})}{L} \le \frac{M - f(x_k)}{L} \operatorname{Max}\left\{1 - \frac{M - f(x_k)}{L}, \frac{1}{2}\right\}, \tag{6.9}$$

which is the same as (4.9) with $\frac{\widetilde{W}W}{4L}$ replaced by $\frac{M - f}{L}$, so that $\frac{M - f(x_k)}{L}$ converges to zero like $\frac{1}{k}$, and $f(x_k) \to M$.

Finally, we have in analogy to the Approximation Criterion of Section 3 the estimate for M given by (6.7):

$$f(x_k) \le M \le f(x_k) + \partial f(x_k)(z_{k+1} - x_k). \tag{6.10}$$

This estimate improves throughout the process; letting $z_{k+1} - x_k = y_k$, we show $\partial f(x_k) y_k \to 0$.

Let $\delta_k = \operatorname{Min}\{\partial f(x_k) y_k, L\}$. Then for some w between x and $x + \frac{\delta_k}{L} y$,

$$M \ge f(x + \frac{\delta_k}{L} y_k) = f(x_k) + \frac{\delta_k}{L} \partial f(x_k) y_k + \frac{1}{2} \frac{\delta_k^2}{L^2} y_k' \partial^2 f(w) y_k$$

$$\ge f(x_k) + \frac{\delta_k}{L} \partial f(x_k) y_k - \frac{\delta_k^2}{2L} \ge f(x_k) + \frac{\delta_k^2}{2L}.$$

Since $f(x_k) \to M$, we have $\delta_k \to 0$.

In the absence of the property (A), a modification of the gradient-and-interpolation procedure which takes account of possible infinite solutions of the linear program (6.4) can be made, for which convergence of the function to the maximum can be shown; however, estimates of the rate of convergence (6.9) and error (6.10) cannot be constructed.

7. APPENDIX

(a) The duality theorem for linear programming asserts, where A, b, c are, respectively, $m \times n$, $m \times 1$, $1 \times n$ matrices:

$$\sup\{cx \mid x \ge 0,\ Ax \le b\} = \inf\{ub \mid u \ge 0,\ uA \ge c\},$$

the supremum (infimum) over an empty set being $-\infty$ (∞); the extrema are assumed if finite.

(b) The simplex method for solving the linear program $\operatorname{Max}\{cx \mid x \ge 0,\ Ax = b\}$ employs at each step the basis consisting of m columns of A such that their totality B is nonsingular and $B^{-1} b \ge 0$. The vector having components $B^{-1} b$ with appropriate zeroes adjoined is a basic feasible vector. A change of basis replaces one column of B by another column from A in such a way as to increase the objective cx on the basic feasible vector. This is continued until the maximum is attained. The constraints are non-degenerate if always $B^{-1} b > 0$. The non-degeneracy assumption of [3] has been removed [4].

(c) An initial basic feasible vector for the linear program is found by applying the simplex method to the expanded problem $\operatorname{Max}\{-y \mid x \ge 0,\ y \ge 0,\ Ax + Iy = b\}$, $x = 0$, $y = b \ge 0$

being an initial basic feasible vector for this problem. If Max < 0, then the original problem is not feasible; otherwise, for the Max, $y = 0$, so x is a basic feasible vector. This formulation is used with a further modification to eliminate degeneracy [2, 4].

(d) The function f (having continuous second derivative) on the convex set S is concave if, and only if, for all x, y in S and $0 \leq \mu \leq 1$, using $h(\mu) = f(x + \mu (y - x))$, we have $h''(\mu) \leq 0$, whence:

$$(y - x)' \partial^2 f(w) (y - x) \leq 0 , \tag{7.1}$$

for all w between x and y, is a necessary and sufficient condition for concavity. If S has interior points, (7.1) is equivalent to the condition that the matrix $\partial^2 f(w)$ be negative semi-definite.

If $f(x) = px - x'Cx$ then $\partial^2 f = -C$, so that f is concave for all x if, and only if, C is positive definite.

(e) For f as in (d), if $x, y \in S$, by Taylor's theorem

$$f(y) - f(x) = \partial f(x) (y - x) + \frac{1}{2} (y - x)' \partial^2 f(w) (y - x)$$

for some w between x and y, so that

$$f(y) - f(x) \leq \partial f(x) (y - x) \tag{7.2}$$

is a necessary condition for concavity of f [6, Lemma 3]. If the left of (7.1) is positive for some x, y, w, then $\overline{x}, \overline{y}$ can be chosen sufficiently close to w that $(\overline{y} - \overline{x})' \partial^2 f(\overline{w}) (\overline{y} - \overline{x}) > 0$ for all $\overline{w}$ between $\overline{x}$ and $\overline{y}$, so that (7.2) fails, whereby it is also sufficient.

(f) Since, in general, if f is concave then $\{x \mid f(x) \geq c\}$ is convex for all c, the solution set $\{x \in S \mid f(x) \geq \text{Max}\}$ of any concave program on a convex set S is convex.

(g) If C is positive semi-definite and symmetric and $x'Cx = 0$ then for all μ and z, $0 \leq (x + \mu z)' C (x + \mu z) = 2\mu x'Cz + \mu^2 z'Cz$, so that $x'Cz = 0$ for all z, whence $x'C = Cx = 0$.

(h) If x and y belong to the solution set of the quadratic problem, then for all $0 \leq \mu \leq 1$,

$$M = f(x + \mu (y - x)) = f(x) + \mu \partial f(x) (y - x) - \frac{1}{2}\mu^2 (y - x)' C (y - x) ,$$

whence $(y - x)' C (x - y) = 0$, so $(y - x)' C = 0$, hence $f(x) = p - 2x'C = p - 2y'C = f(y)$. Thus, ∂f is constant on the solution set. Furthermore, $f(x + \mu (y - x)) = f(x)$ for all μ, so that the solution set is the intersection of the constraint set with some linear manifold.

(i) We prove by induction on k that if the quadratic function $f(x) = px + x'Cx$ (not necessarily concave) is bounded on the polyhedral convex set R of dimension k, then it assumes its supremum on R.

We may write any $k+1$-dimensional polyhedron R as $\{s + \mu t \mid s \in S, t \in T, \mu \geq 0\}$, where S is a certain bounded convex polyhedron and T is the intersection of a certain convex polyhedral cone with the unit sphere. Since for $r \in R$, $t \in T$, $f(r + \mu t) = f(x) + \mu (p + 2r'C)t + \mu^2 t'Ct$ is bounded for $\mu \geq 0$, $t'Ct \leq 0$ for all t in T.

If, on the one hand, $t'Ct < 0$ for all t in T, then there exist $\delta > 0$ and D such that $t'Ct < -\delta$ and $(p + 2s'C)t < D$ for all t in T and s in S, so that the maximum of f is assumed on the compact set $S + \frac{D}{\delta} T$.

Suppose, on the other hand, some $t_0' C t_0 = 0$. If for all $r \in R$ we have $r + \mu t_0 \in R$ for all μ, then the boundedness of $f(r + \mu t)$ implies $(p + 2r'C)t_0 = 0$ for all r, so that the values of f on R are unchanged by projection into the k-dimensional subspace normal to t_0, to which the induction hypothesis may be applied.

Otherwise, for each r, $r + \mu t_0 \notin R$ for some μ, so that for each r, $b_r = r + \text{Min}\{\mu \mid r + \mu t_0 \in R\} t_0$ lies on the boundary of R, and $f(b_r) \geq f(r)$, since $(p + 2r'C)t_0 \leq 0$. Since the supremum of f on each k-dimensional bounding hyperplane of R is assumed, so is it on R. (Irving Kaplansky has pointed out that the above result does not obtain for polynomials of degree greater than two. It is clearly not true for polynomials of odd degree; and the function $x^2 + (1 - xy)^2$ does not assume its greatest lower bound, zero, on the plane.)

REFERENCES

[1] Barankin, E. W., and Dorfman, R., Toward Quadratic Programming, O. N. R. Logistics Projects at Columbia Univ. and U. C. L. A., Univ. of California at Berkeley.

[2] Charnes, A., Cooper, W. W., and Henderson, A., An Introduction to Linear Programming, Wiley, 1953.

[3] Dantzig, G. B., "Maximization of a linear function of variables subject to linear inequalities," Activity Analysis of Production and Allocation (T. C. Koopmans ed.), Wiley, 1951.

[4] Dantzig, G. B., Orden, A., and Wolfe, P., "The generalized simplex method for minimizing a linear form under linear inequality restraints," Pacific J. Math. 5, 183-195 (June 1955).

[5] König, D., Theorie der Endlichen und Unendlichen Graphen, Leipzig, 1936.

[6] Kuhn, H. W., and Tucker, A. W., "Nonlinear programming," Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability, 481-492 (1951).

[7] Saaty, T. L., "The number of vertices of a polyhedron," Amer. Math. Monthly 62, 326-331 (May 1955).

ADDITIONAL BIBLIOGRAPHY

(Items 4-7 discuss computational proposals.)

1. Debreu, G., "Definite and semidefinite quadratic forms," Econometrica 20, 295-300 (1952).
2. Mann, H. B., Quadratic forms with linear constraints," Amer. Math. Monthly, 58, 430-433 (1943).
3. Phipps, C. G., Maxima and minima under restraint," Amer. Math. Monthly 59, 230-235 (1952).
4. Arrow, K. J., and Hurwicz, L., "A gradient method for approximating saddle points and constrained maxima," RAND P-223 (June 1951). To appear in Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.
5. Charnes, A., and Lemke, C. E., "Minimization of non-linear separable convex functionals (Computational theory of linear programming IV)," O.N.R. Research Memorandum No. 16, Carnegie Institute of Technology (May, 1954).

391368 O - 56 - 8

6. Manne, A. S., "Concave programming for gasoline blends," RAND P-383 (April 1953).
7. Markowitz, H., "The optimization of quadratic functions subject to linear constraints," Naval Research Logistics Quarterly, Vol. 3, Nos. 1, 2, 1956.

* * *

THE OPTIMIZATION OF A QUADRATIC FUNCTION SUBJECT TO LINEAR CONSTRAINTS

Harry Markowitz[1]

The author discusses a computational technique applicable to the determination of the set of "efficient points" for quadratic programming problems.

1. QUADRATIC PROBLEMS

Suppose that variables $X_1, \ldots, X_N$ are to be chosen subject to linear constraints:

$$\Sigma a_{ij} X_j = b_i, \quad i = 1, \ldots, m_1 \tag{1}$$

$$\Sigma a_{ij} X_j \geq b_i, \quad i = m_1 + 1, \ldots, m \tag{2}$$

$$X_j \geq 0, \quad j = 1, \ldots, N_1 \tag{3}$$

where $0 \leq m_1 \leq m$, $0 \leq N_1 \leq N$ and the matrix (a_{ij}) $i = 1, \ldots, m_1$ has rank m_1 (otherwise the system is inconsistent or has at least one redundant equation). The payoff is a linear function $R = \Sigma r_j X_j$ whose coefficients r_j are not known at the time the X_j are chosen. The r_j, rather, are random variables with expected values μ_j and covariances σ_{jk} (including variances $\sigma_{jj} = \sigma_j^2$). The expected value of R is

$$E = \Sigma \mu_j X_j. \tag{4}$$

The variance of R is

$$V = \Sigma\Sigma \sigma_{jk} X_j X_k. \tag{5}$$

Suppose further that some decision-maker likes expected payoff (E) and dislikes variance of payoff (V). Our problem is to compute for the decision-maker (a) the "efficient combinations" of E and V, i.e., those attainable (E, V) combinations which give minimum V for given E and maximum E for given V (Figure 1); and (b) the points in the X space associated with the efficient E, V combinations, i.e., the set of efficient X's.

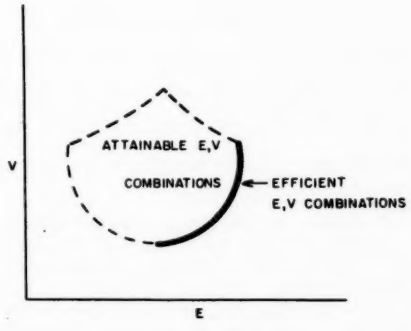


Figure 1

[1]The writer has particularly benefited from discussions with Kenneth Arrow on the subject matter of this paper.

A computing technique is presented in this paper for generating the above efficient sets. An adaptation of this technique can be used for problems of maximizing or minimizing quadratic forms (with the "right" properties) subject to linear constraints.

The practical problem which first suggested the above computing problem was that of selecting a portfolio of securities.[2] Here X_j is the amount invested in the j^{th} security; the μ_j and σ_{jk} are the expected returns and covariances of returns from the various securities. In the simplest case the constraint set is $\Sigma X_j = 1$, $X_j \geq 0$. A problem of very similar structure, analyzed independently by H. S. Houthakker,[3] is that of finding the expenditure on various goods as a function of income for an individual whose utility function is of the form $u = \Sigma a_j X_j + \Sigma\Sigma a_{ij} X_i X_j$. A problem of maximizing a monopolist's quadratic profit function subject to linear constraints is presented by Robert Dorfman.[4] Another problem of this general character is that of maximizing a quadratic likelihood function where there is a priori information concerning the values of parameters to be estimated. Now that reasonably convenient computing procedures exist for such quadratic problems we may be permitted the hope that still other classes of interesting questions can be reduced to this form.

This paper will discuss only minimization problems involving the quadratic form $\Sigma\Sigma\sigma_{ij} X_i X_j$ whose matrix (σ_{ij}) is positive semi-definite. The reader should have no difficulty in extending the results to minimization problems involving $\Sigma\rho_j X_j + \Sigma\Sigma\sigma_{ij} X_i X_j$ where (σ_{ij}) is positive semi-definite or maximization problems where (σ_{ij}) is negative semi-definite.

2. ASSUMPTIONS

According to customary usage we say:

(a) A set of points (S) (in Euclidean n-space) is convex if $X^{(1)} \in S$ and $X^{(2)} \in S$ imply $\lambda X^{(1)} + (1-\lambda) X^{(2)} \in S$, for any $0 \leq \lambda \leq 1$.

(b) A set is closed if $X_1, \ldots, X_t, \ldots \longrightarrow y$ and $X_1, \ldots, X_t, \ldots \in S$ imply $y \in S$.

(c) A function $f(X)$ is convex over a set S if $X^{(1)} \in S$, $X^{(2)} \in S$ and $[\lambda X^{(1)} + (1-\lambda) X^{(2)}] \in S$ imply $f(\lambda X^{(1)} + (1-\lambda) X^{(2)}) \leq \lambda f(X^{(1)}) + (1-\lambda) f(X^{(2)})$ for all $0 \leq \lambda \leq 1$.

(d) A function is strictly convex over a set S if $X^{(1)} \in S$, $X^{(2)} \in S$ and $[\lambda X^{(1)} + (1-\lambda) X^{(2)}] \in S$ imply $f(\lambda X^{(1)} + (1-\lambda) X^{(2)}) < \lambda f(X^{(1)}) + (1-\lambda) f(X^{(2)})$ for all $0 < \lambda < 1$.

The set $(\bar{S})$ of points which satisfy constraints (1), (2), and (3) is a closed, convex set. Variance (V) is a positive semi-definite quadratic form, i.e., $\sum_j \sum_k \sigma_{jk} X_j X_k \geq 0$ for all $(X_1, \ldots, X_N)$. It is also convex. The covariance matrix (σ_{jk}) is non-singular if, and only if, V is positive definite $\left\{\text{i.e., } \sum_j \sum_k \sigma_{jk} X_j X_k > 0, \text{ if } (X_1, \ldots, X_N) \neq (0, \ldots, 0)\right\}$, which in turn is true if, and only if, V is strictly convex over the set of all X.[5]

---

[2] Harry Markowitz, "Portfolio Selection," Journal of Finance, 1952.

[3] "La Forme Des Courbes D'Engel," Cahiers du Séminaire d'Econometrie, 1953.

[4] Application of Linear Programming to the Theory of the Firm, University of California Press, 1951.

[5] That $V(X) \geq 0$, for all X, is due to the fact that V is the expected value of a square, $E(r - Er)^2$, and therefore cannot be negative. That $|\sigma_{ij}| \neq 0$ if, and only if, V is positive definite is a corollary of material found, e.g., in Birkhoff and MacLane, A Survey of Modern Algebra, Chapter IX, particularly section 8, pp. 243-247. The implications of positive definiteness and semi-definiteness for convexity may be seen as follows: Let $C = (\sigma_{ij})$. Let X and Y be column vectors; X' and Y' be row vectors. C is symmetric so that $C = C'$ and $X'CY = Y'CX$ for any X, Y. We wish to see the implications of

(1) $X'CX \geq 0$

(2) $X'CX = 0 \Longleftrightarrow X = 0$

(Continued)

We will assume[19] that $\bar{S}$ is not vacuous. We will also assume that V is strictly convex over the set of X's which satisfy the equations

$$\Sigma a_{ij} X_j = b_i , \quad i = 1, \ldots, m_1.$$

This assures[6] us that V takes on a unique minimum over $\bar{S}$ and that if $E = E^0$ is attainable in $\bar{S}$, then V takes on unique minimum over the set

$$\bar{S} \cap \{X \mid \Sigma \mu_j X_j \geq E_0\}.$$

If a function is convex over a set S, it is convex over any subset of S; therefore $|\sigma_{ij}| \neq 0$ implies that V is strictly convex over $\{X \mid \Sigma a_{ij} X_j = b_i, i = 1, \ldots, m_1\}$. This is not a necessary condition, however. Necessary and sufficient conditions on A and $(\sigma_{ij}) = C$ are discussed in the footnote.[7]

---

[5](Continued)
for the difference

$$D = [\lambda X'CX + (1-\lambda) Y'CY] - [(\lambda X' + (1-\lambda)Y') C(\lambda X + (1-\lambda)Y)].$$

Expanding the second term and subtracting we get

$$D = \lambda(1-\lambda) \cdot [X'CX - 2X'CY + Y'CY]$$
$$= \lambda(1-\lambda) \cdot [(X'-Y') C (X-Y)].$$

Assumption (1) implies $D \geq 0$ for all X,Y. Assumptions (1) and (2) imply $D > 0$ if $X \neq Y$. Conversely, if D is positive for all $X \neq Y$, letting $Y = 0$ we find $X'CX > 0$ for all $X \neq 0$.

[6]Since equations (1) have rank m_1, m_1 variables and the m_1 equations could be eliminated (as in footnote 7) to leave a system with $N-m_1$ variables and $(m-m_1) + N_1$ inequalities. V is strictly convex in these $N-m_1$ variables and therefore the associated quadratic is positive definite. Let Y be any point in the space of the $N-m_1$ variables satisfying the $(m-m_1) + N_1$ inequalities. The points which satisfy these inequalities and have $V \leq V(Y)$ form a compact, convex set. Since V is continuous, it attains its minimum at least once on this set. Since it is strictly convex, it attains its minimum only once. The same argument applies if the constraint

$$\Sigma \mu_j X_j \geq E^0$$

is added to the $(m-m_1) + N_1$ inequalities.

[7]Suppose $m_1 \geq 1$. Since $\begin{pmatrix} a_{11} \cdots a_{1N} \\ \vdots \quad \vdots \\ a_{m_1 1} \cdots a_{m_1 N} \end{pmatrix}$ has rank m_1 we can write, after perhaps relabeling variables,

$$\begin{pmatrix} a_{11} \cdots a_{1m_1} \\ \vdots \quad \vdots \\ a_{m_1 1} \cdots a_{m_1 m_1} \end{pmatrix} \begin{pmatrix} X_1 \\ \vdots \\ X_{m_1} \end{pmatrix} + \begin{pmatrix} a_{1m_1+1} \cdots a_{1N} \\ \vdots \quad \vdots \\ a_{m_1 m_1+1} \cdots a_{m_1 N} \end{pmatrix} \begin{pmatrix} X_{m_1+1} \\ \vdots \\ X_N \end{pmatrix} = \begin{pmatrix} b_1 \\ \vdots \\ b_{m_1} \end{pmatrix}$$

or $A^{(1)} X^{(1)} + A^{(2)} X^{(2)} = b$ where $A^{(1)}$ is non-singular. We thus have $X^{(1)} = (A^{(1)})^{-1} b - (A^{(1)})^{-1} A^{(2)} X^{(2)}$. We can express V in terms of $X^{(2)}$ by substitution, i.e.,

$$X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix} = \begin{pmatrix} A^{(1)-1} b \\ 0 \end{pmatrix} - \begin{pmatrix} A^{(1)-1} A^{(2)} \\ I \end{pmatrix} X^{(2)}$$

$$V = X'CX = (X^{(1)'} X^{(2)'}) C \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$$

(Continued)

3. THE CRITICAL LINE $\bar{\lambda}$

Let us first note the answer to a simpler problem than that of finding all (E, V) efficient points. Suppose we wished to minimize V subject to the equations

$$\Sigma a_{ij} X_j = b_i , \quad i = 1, \ldots, m_1$$

without regard to the inequalities (2) and (3). A necessary condition for a minimum is that $X_1, \ldots, X_N$ be a solution to the Lagrangian equations,

$$\frac{\partial \left(\Sigma\Sigma \sigma_{ij} X_i X_j - 2 \Sigma\lambda_i \Sigma_j a_{ij} X_j\right)}{\partial X_k} = 0 , \quad k = 1, \ldots, N \tag{6}$$

as well as $\Sigma a_{ij} X_j = b_i$, $i = 1, \ldots, m_1$, i.e.,

$$\sum_{j=1}^{N} \sigma_{kj} X_j + \Sigma(-\lambda_i) \alpha_{ik} = 0 , \quad k = 1 \ldots, N \tag{7}$$

$$\Sigma a_{ij} X_j = b_i , \quad i = 1, \ldots, m_1 . \tag{8}$$

Given the assumption that V is strictly convex over $\{X | \Sigma a_{ij} X_j = b_i, \ i = 1, \ldots, m_1\}$ it follows that

$$\begin{pmatrix} \sigma_{11} & \cdots & \sigma_{1N} & a_{11} & \cdots & a_{m_1 1} \\ \vdots & & \vdots & \vdots & & \vdots \\ \sigma_{N1} & \cdots & \sigma_{NN} & a_{1N} & \cdots & a_{m_1 N} \\ a_{11} & \cdots & a_{1N} & 0 & \cdots & 0 \\ \vdots & & & \vdots & & \vdots \\ a_{m_1 1} & \cdots & a_{m_1 N} & 0 & \cdots & 0 \end{pmatrix}$$

---

[7] (Continued)

$$= [(b' A^{(1)'-1}, 0) - (X^{(2)'} A^{(2)'} A^{(1)'-1}, X^{2'})] C \left[\begin{pmatrix} A^{(1)-1} b \\ 0 \end{pmatrix} - \begin{pmatrix} A^{(1)-1} A^{(2)} X^{(2)} \\ X^{(2)} \end{pmatrix}\right]$$

$$= (b' A^{(1)'-1}, 0) C \begin{pmatrix} A^{(1)-1} b \\ 0 \end{pmatrix} - 2X^{(2)} (A^{(2)'} A^{(1)'-1}, I) C \begin{pmatrix} A^{(1)-1} b \\ 0 \end{pmatrix}$$

$$+ X^{(2)'} \left[(A^{(2)'} A^{(1)-1}, I) C \begin{pmatrix} A^{(1)-1} A^{(2)} \\ I \end{pmatrix}\right] X^{(2)} .$$

V is strictly convex for all $X^{(2)}$ if, and only if, the last (i.e., the quadratic) term is strictly convex. This is so if, and only if,

$$(A^{(2)'} A^{(1)-1'}, I) C \begin{pmatrix} A^{(1)-1} A^{(2)} \\ I \end{pmatrix}$$

is non-singular.

is non-singular.[8] Since a strictly convex V takes on a unique minimum on a convex set, the unique solution to (7) and (8) is this minimum.

Next consider the problem of minimizing V subject not only to (1) but also to the constraint that $E = E_0$, i.e.,

$$\sum_{j=1}^{N} \mu_j X_j = E_0 . \tag{9}$$

We must distinguish two cases:

Case 1: The row vector $(\mu_1, \ldots, \mu_N)$ can be expressed as a linear combination of the $(a_{i1}, \ldots, a_{iN})$, i.e., there exists $\alpha_1, \ldots, \alpha_{m_1}$ such that

$$(\alpha_1, \ldots, \alpha_{m_1}) \begin{pmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & & \vdots \\ a_{m_1 1} & \cdots & a_{m_1 N} \end{pmatrix} = (\mu_1, \ldots, \mu_N) . \tag{10}$$

Case 2: There does not exist such a linear combination.

As is shown below,[9] in Case 1 only one value of E, say $E = E^*$, is attainable. Therefore if we require $E \neq E^*$ no solution can be found; if we require $E = E^*$, equations (7) and (8) give the minimum. In Case 2 the matrix

$$\begin{pmatrix} \sigma_{11} & \cdots & 1n a_{11} & \ldots & a_{m_1 N} & \mu_1 \\ \vdots & & \vdots & & & \vdots \\ \sigma_{N1} & \cdots & \sigma_{NN} a_{1N} & \cdots & a_{m_1 N} & \mu_N \\ a_{11} & \cdots & a_{m_1 1} & O \ldots & O & O \\ \vdots & & \vdots & & \vdots & \vdots \\ a_{m_1} & \cdots & a_{m_1 N} & O \ldots & O & O \\ \mu_1 & \cdots & \mu_N & O \ldots & O & O \end{pmatrix}$$

---

[8] If $m_1 = 0$ the statement reduces to one proved in footnote 5. Suppose $m_1 \geq 1$. If $\begin{pmatrix} C & A' \\ A & O \end{pmatrix}$ is singular there is a vector $\begin{pmatrix} Y \\ -\lambda \end{pmatrix} \neq \begin{pmatrix} O \\ O \end{pmatrix}$ such that $\begin{pmatrix} C & A' \\ A & O \end{pmatrix} \begin{pmatrix} Y \\ -\lambda \end{pmatrix} = \begin{pmatrix} O \\ O \end{pmatrix}$ i.e., $\begin{pmatrix} CY \\ AY \end{pmatrix} = \begin{pmatrix} A'\lambda \\ O \end{pmatrix}$. Since the rank of A is m_1 there is no $\lambda \neq O$ such that $A'\lambda = O$, therefore $Y \neq O$ in $\begin{pmatrix} Y \\ -\lambda \end{pmatrix}$ above. $V(Y) = Y'CY = Y'A'\lambda = (AY)'\lambda = O$. Let X be any point in $S = \{X \mid AX = b\}$ where

$X' = (X_1, \ldots, X_N)$

$b' = (b_1, \ldots, b_{m_1})$

$A(X + Y) = AX + O = b$; therefore $X + Y$ is in S

$V(X) = X'CX$

$V(X + Y) = X'CX + 2X'CY = X'CX + 2X'A'\lambda = X'CX + 2b'\lambda$

$V(1/2X + 1/2(X + Y)) = V(X + 1/2\,Y)$

$= X'CX + X'CY$

$= X'CX + b'\lambda$

$= 1/2\,V(X) + 1/2\,V(X+Y)$,

thus contradicting strict convexity.

[9] If $\alpha'A = \mu'$ and $AX = b$ then $E = \mu'X = \alpha'AX = \alpha'b$.

is non-singular,[10] and therefore the equations

$$\Sigma\sigma_{jk} X_k + \Sigma(-\lambda_1) a_{ij} - \lambda_E \mu_j = 0 \quad , \quad j = 1, \ldots, N \tag{11}$$

$$\Sigma a_{ij} X_j = b_i \quad , \quad i = 1, \ldots, m_1 \tag{12}$$

$$\Sigma \mu_j X_j = E^0 \tag{13}$$

have a unique solution which gives minimum V for the specified E^0. If we let E^0 go from $-\infty$ to $+\infty$, the solution to (11), (12), and (13) traces out a line in the (X,λ) space. This line may also be described as the solution to the following $N + m_1$ equations in $N + m_1 + 1$ unknowns:

$$\sum_{j=1}^{N} \sigma_{jk} X_k - \sum_{i=1}^{m_1} \lambda_i a_{ij} - \lambda_E \mu_j = 0 \quad , \quad j = 1, \ldots, N \tag{14}$$

$$\Sigma a_{ij} X_j = b_i \quad , \quad i = 1, \ldots, m_1 \tag{15}$$

or

$$\Sigma\sigma_{jk} X_k + \Sigma(-\lambda_j) a_{ij} = \lambda_E \mu_j \quad , \quad j = 1, \ldots, N \tag{16}$$

$$\Sigma \alpha_{ij} X_j = b_i \quad , \quad i = 1, \ldots, m_1 \tag{17}$$

$$\text{for } -\infty \leq \lambda_E \leq +\infty .$$

Since the matrix of equations (16) and (17) is non-singular, given our assumption of strict convexity, they have a solution for every value of λ_E whether we are in Case 1 or Case 2 above. In Case 1 the values of $X_1, \ldots, X_N$ do not change (only the values of the λ's change) as λ_E goes from $-\infty$ to $+\infty$.[11] In Case 2 the X's as well as the λ's change. In Case 2 we can define $\hat{V}(E)$ to be minimum V as a function of E. $2\lambda_E = d\hat{V}/dE \cdot \hat{V}(E)$ must be strictly convex; therefore, E increases with λ_E. In Section 11 we show that $\hat{V}(E)$ is a parabola.

4. CRITICAL LINES $\ell(\bar{s},\bar{\imath})$

The set of points (X,λ) which satisfy (16) and (17) will be referred to as the critical line $\bar{\ell}$ associated with the subspace

$$\bar{S} = \{X \mid \Sigma a_{ij} X_j = b_i \text{ for } i = 1, \ldots, m_1\}.$$

Critical lines will also be associated with certain other subspaces.

---

[10] Same proof as in footnote 8, using the fact that $\begin{pmatrix} A \\ \mu' \end{pmatrix}$ has rank $m_1 + 1$.

[11] For if $\mu = A'\tilde{\lambda}$ and $\begin{pmatrix} C & A' \\ A & O \end{pmatrix} \begin{pmatrix} X^0 \\ \lambda^0 \end{pmatrix} = R + \begin{pmatrix} \mu \\ O \end{pmatrix} \lambda^0_E$ then $\begin{pmatrix} C & A' \\ A & O \end{pmatrix} \begin{pmatrix} X^0 \\ \lambda^0 + \tilde{\lambda}\theta \end{pmatrix}$ $= R + \begin{pmatrix} \mu \\ O \end{pmatrix} \left(\lambda^0_E + \theta \right)$.

Let $X_{j_1}, \ldots, X_{j_J}$ be a subset of variables. Let

$$\Sigma a_{ij} X_j = b_i \ , \quad i = i_1, \ldots, i_I$$

be a subset of the constraints (1) and (2) with the inequalities replaced by equalities when $i > m_1$. Let $\mathcal{I}$ be the ordered set of indices $(i_1, \ldots, i_I)$; let $\mathcal{J}$ be the ordered set $(j_1, \ldots, j_J)$. We will be particularly interested in $\mathcal{I}$ and $\mathcal{J}$ of the form

(18) $$\mathcal{I} = \{1, 2, \ldots, m_1, i_{m_1+1}, \ldots, i_I\} \quad \text{where } I \geq m_1$$

(19) $$\mathcal{J} = \{j_1, \ldots, j_L, N_1 + 1, \ldots, N\} \quad \text{where } 0 \leq L \leq N_1 .$$

For any indices $\mathcal{I}$ and $\mathcal{J}$ satisfying (18) and (19) we define the submatrix

(20) $$A_{\mathcal{I}\mathcal{J}} = \begin{pmatrix} a_{i_1 j_1} & \cdots & a_{i_1 j_J} \\ \vdots & & \vdots \\ a_{i_I j_1} & \cdots & a_{i_I j_J} \end{pmatrix}.$$

We similarly define subvectors $X_{\mathcal{J}}$, $\lambda_{\mathcal{I}}$ and

(21) $$B_{\mathcal{I}} = \begin{pmatrix} b_{i_1} \\ \vdots \\ b_{i_I} \end{pmatrix};$$

also submatrices

(22) $$C_{\mathcal{J}\mathcal{J}} = \begin{pmatrix} \sigma_{j_1 j_1} & \cdots & \sigma_{j_1 j_J} \\ \vdots & & \vdots \\ \sigma_{j_J j_1} & \cdots & \sigma_{j_J j_J} \end{pmatrix}$$

(23) $$M_{\mathcal{I}\mathcal{J}} = \begin{pmatrix} C_{\mathcal{J}\mathcal{J}} & A'_{\mathcal{I}\mathcal{J}} \\ A_{\mathcal{I}\mathcal{J}} & O \end{pmatrix}.$$

If $I = m_1 = 0$, $\mathcal{I}$ is empty. In this case it will sometimes be convenient to think of $A_{\mathcal{I}\mathcal{J}}$ as having no rows and J columns. To every $(\mathcal{I}, \mathcal{J})$ satisfying (18) and (19) we associate a subspace

$$S(\mathcal{I}, \mathcal{J}) = \{X \mid X_j = 0 \text{ for } j \notin \mathcal{J},\ A_{\mathcal{I}\mathcal{J}} X_{\mathcal{J}} = B_{\mathcal{I}}\}.$$

If $A_{\mathcal{I}\mathcal{J}}$ has no rows this reduces to $S(\mathcal{J}) = \{X \mid X_j = 0 \text{ for } j \notin \mathcal{J}\}$. Since $\mathcal{I}$ and $\mathcal{J}$ satisfy (18) and (19), $S(\mathcal{I}, \mathcal{J}) \subset \bar{S}$. Since V is strictly convex over $\bar{S}$, it is strictly convex over $S(\mathcal{I}, \mathcal{J})$.

$A_{\mathcal{I}\mathcal{J}}$ has a rank of I or less. If $A_{\mathcal{I}\mathcal{J}}$ has rank I then the matrix

$$M_{\mathcal{I}\mathcal{J}} = \begin{pmatrix} C_{\mathcal{J}\mathcal{J}} & A_{\mathcal{I}\mathcal{J}} \\ A_{\mathcal{I}\mathcal{J}} & O \end{pmatrix} \tag{24}$$

is non-singular. (This is a special case of the proposition proved in footnote 8.) If $A_{\mathcal{I}\mathcal{J}}$ has rank less than I, $M_{\mathcal{I}\mathcal{J}}$ is singular, for its last I rows are not independent. For every $(\mathcal{I}, \mathcal{J})$ satisfying (18) and (19) whose $A_{\mathcal{I}\mathcal{J}}$ has rank equal to the number of its rows, we define the critical line $\ell_{\mathcal{I}\mathcal{J}}$ to be the set of points $(X_1, \ldots, X_N, \lambda_1, \ldots, \lambda_m)$ which satisfy

$$\begin{aligned} X_j &= 0 \quad \text{for } j \notin \mathcal{J} \\ \lambda_i &= 0 \quad \text{for } i \notin \mathcal{I} \end{aligned} \tag{25}$$

and

$$\begin{pmatrix} X_{\mathcal{J}} \\ -\lambda_{\mathcal{I}} \end{pmatrix} = M^{-1}_{\mathcal{I}\mathcal{J}} \begin{pmatrix} O \\ B_{\mathcal{I}} \end{pmatrix} + M^{-1}_{\mathcal{I}\mathcal{J}} \begin{pmatrix} \mu_{\mathcal{J}} \\ O \end{pmatrix} \lambda_E \, .$$

Equations (25) may be written in the form

$$\left.\begin{aligned} X_j &= \alpha_{Xj} + \beta_{Xj} \lambda_E \qquad (26) \\ \lambda_i &= \alpha_{\lambda i} + \beta_{\lambda i} \lambda_E \qquad (27) \end{aligned}\right\} -\infty < \lambda_E < \infty \, .$$

Equations (26) by themselves are the projection of $\ell(\mathcal{I}, \mathcal{J})$ onto the X-space. As with $\bar{S}$ and $\bar{\ell}$ we have two cases:

(1) Only one value of E is obtainable in $S(\mathcal{I}, \mathcal{J})$ and the X-projection is a point.

(2) All values of E are obtainable and the X-projection is a line. This line is the set of X's in $S(\mathcal{I}, \mathcal{J})$ which give minimum V for some E. Let

$$\xi_i = \Sigma a_{ij} X_j - b_i \quad , \quad i = 1, \ldots, m \tag{28}$$

$$\eta_j = 1/2 \, \frac{\partial V - 2 \sum_{i=1}^{m} \lambda_i \Sigma a_{ij} X_j - 2\lambda_E \Sigma \mu_j}{\partial X_j} \tag{29}$$

$$= \sum_{j=1}^{N} \sigma_{jk} X_k + \sum_{i} (-\lambda_l) a_{ij} - \mu_j \lambda_E \, .$$

Constraints (1) and (2) state that

$$\xi_i = 0 \quad \text{for } i = 1, \ldots, m_1$$

$$\xi_i \geq 0 \quad \text{for } i = m_1 + 1, \ldots, m.$$

Along any critical line we have

$$\begin{aligned} X_j &= 0 \quad \text{for } j \notin \mathcal{J}, \\ \eta_j &= 0 \quad \text{for } j \in \mathcal{J}, \\ \xi_i &= 0 \quad \text{for } i \in \mathcal{I}, \\ \lambda_i &= 0 \quad \text{for } i \notin \mathcal{I}. \end{aligned} \tag{30}$$

Also, from (25), letting m^{st} be the $(s,t)^{th}$ element of $M_{\mathcal{I}\mathcal{J}}^{-1}$, we have

$$\begin{aligned} X_{j_s} &= \sum_{h=1}^{I} m^{s,h+J} b_{i_h} + \left(\sum_{h=1}^{J} m^{sh} \mu_{j_h} \right) \lambda_E \\ &= \alpha_{Xj_s} + \beta_{Xj_s} \lambda_E \qquad \text{for } s = 1, \ldots, J. \end{aligned} \tag{31}$$

$$\begin{aligned} \lambda_{i_s} &= -\sum_{h=1}^{I} m^{s+J,\,h+J} b_{i_h} - \left(\sum_{h=1}^{J} m^{s+J,\,h} \mu_{j_h} \right) \lambda_E \\ &= \alpha_{\lambda i_s} + \beta_{\lambda i_s} \lambda_E \qquad \text{for } s = 1, \ldots, I. \end{aligned} \tag{32}$$

From (28) and (29) we have

$$\begin{aligned} \xi_i &= \sum_{h=1}^{J} a_{ij_h} \alpha_{Xj_h} - b_i + \left(\sum_{h=1}^{J} a_{ij_h} \beta_{Xj_h} \right) \lambda_E \\ &= \alpha_{\xi_i} + \beta_{\xi_i} \lambda_E ; \end{aligned} \tag{33}$$

$$\begin{aligned} \eta_j &= \left(\sum_{h=1}^{J} \sigma_{jj_h} \alpha_{Xj_h} - \Sigma a_{i_h j} \alpha_{\lambda i_h} \right) \\ &\quad + \left(\sum_{h=1}^{J} \sigma_{jj_h} \beta_{Xj_h} - \sum_{h=1}^{I} a_{i_h j} \beta_{\lambda i_h} - \mu_j \right) \lambda_E \\ &= \alpha_{\eta j} + \beta_{\eta j} \lambda_E . \end{aligned} \tag{34}$$

A corollary of the results of an important paper by Kuhn and Tucker[12] is that a sufficient condition for a point X to give minimum V for a set

$$\bar{S} \cap \{X \mid \Sigma \mu_j X_j \geq E_0\}$$

for some E_0 is that

$$\begin{aligned} X_j &\geq 0 \quad \text{for } j \leq N_1 \text{ and in } \mathcal{J}, \\ \eta_j &\geq 0 \quad \text{for } j \notin \mathcal{J}, \\ \xi_i &\geq 0 \quad \text{for } i \notin \mathcal{A}, \\ \lambda_i &\geq 0 \quad \text{for } i > m_1 \text{ and in } \mathcal{A}, \end{aligned} \tag{35}$$

and $\lambda_E \geq 0$. If $\lambda_E > 0$, the constraint $E \geq E_0$ is effective; if E_0 were increased, an equally low value of V could not be obtained. If $\lambda_E = 0$, the point gives minimum V in $\bar{S}$. In either case the point is efficient.

It will be convenient at times to employ the following relabeling of variables:

$$\begin{aligned} v_k &= X_k && \text{for } k = 1, \ldots, N_1, \\ v_k &= \eta_{k-N_1} && \text{for } k = N_1 + 1, \ldots, 2N_1, \\ v_k &= \xi_{m_1 + k - 2N_1} && \text{for } k = 2N_1 + 1, \ldots, 2N_1 + m - m_1, \\ v_k &= \lambda_{k + 2m_1 - m - 2N_1} && \text{for } k = 2N_1 + m - m_1 + 1, \ldots, 2N_1 + 2m - 2m_1. \end{aligned} \tag{36}$$

Also

$$K = 2N_1 + 2m - 2m_1 \tag{37}$$

and

$$\mathcal{K} = \{\text{the set of } k \text{ which identify the variables in equation 30)}\}. \tag{38}$$

Thus on any critical line we have

$$v_k = 0 \quad \text{for } k \in \mathcal{K} \tag{39}$$

and a point X is efficient if it is a projection of a point on a critical line with

$$v_k \geq 0 \quad \text{for } k \notin \mathcal{K}, \tag{40}$$

or

$$v_k \geq 0 \quad \text{for } k = 1, \ldots, K. \tag{41}$$

[12] H. W. Kuhn and A. W. Tucker, "Nonlinear Programming," in Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability.

5. INTERSECTIONS OF CRITICAL LINES; NON-DEGENERACY CONDITIONS

In the computing procedure of the next section we move along a critical line until it intersects a plane $v_k = 0$, $k = 1, \ldots, K$. Then either one row and the corresponding column is added to M, or one row and the corresponding column is deleted from M. This raises two questions: (1) under what conditions will the matrix obtained by such additions or deletions be non-singular, and (2) how should the new inverse be obtained? The latter question will not be discussed except to note that the possession of the old inverse is of great value in obtaining the new one.[13]

Concerning the former question, suppose $M_{\mathcal{S}\mathcal{G}}$, with $\mathcal{S}\mathcal{G}$ satisfying (18) and (19), is non-singular and thus defines a critical line ℓ. Suppose ℓ intersects (but is not contained in) the plane $v_k = 0$, $1 \leq k \leq K$. We distinguish four cases, depending on whether v corresponds to an X, an η, a λ, or a ξ:

1. The deletion of a variable. Suppose ℓ intersects a plane $X_j = 0$, $j = 1, \ldots, N_1$. Suppose that j is deleted from the set $\mathcal{G}$ leaving $\mathcal{G}^*$. Is $M_{\mathcal{S}\mathcal{G}^*}$ non-singular? We may suppose without loss of generality that $j = j_1$ and that $A_{\mathcal{S}\mathcal{G}}$ may therefore be written

$$A_{\mathcal{S}\mathcal{G}} = (\alpha A_{\mathcal{S}\mathcal{G}^*}) \tag{42}$$

where α is the column to be deleted. The matrix $\begin{pmatrix} \alpha & A_{\mathcal{S}\mathcal{G}^*} \\ 1 & 0 \ldots 0 \end{pmatrix}$ has either rank, I or I + 1. If it has rank I, then

$$\begin{pmatrix} C_{\mathcal{S}\mathcal{G}} & A'_{\mathcal{S}\mathcal{G}} & \mu_j \\ & & \vdots \\ & & \mu_{j_J} \\ A_{\mathcal{S}\mathcal{G}} & O & 0 \\ & & \vdots \\ & & 0 \\ 1\ 0 \ldots 0 & 0 \ldots 0 & 0 \end{pmatrix} = \widetilde{M} \tag{43}$$

is singular. In this case the equations

$$\widetilde{M} \begin{pmatrix} X_{\mathcal{G}} \\ \lambda_{\mathcal{S}} \\ \lambda_E \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \\ b_i \\ \vdots \\ b_{i_I} \\ 0 \end{pmatrix} \tag{44}$$

have either no solution or an infinity of solutions. Thus if (44) has one solution, i.e., if ℓ intersects $X_{j_1} = 0$ (but is not contained in it), $\widetilde{M}$ is non-singular and the rank of $\begin{pmatrix} \alpha & A_{\mathcal{S}\mathcal{G}^*} \\ 1\ 0 & 0 \end{pmatrix}$ is I + 1;

---

[13] This involves procedures similar to those used in removing a variable from a regression analysis or modifying a basis in linear programming; e.g., see R. A. Fisher, Statistical Methods for Research Workers, p. 164, 10th ed., and R. Dorfman, Activity Analysis of Production and Allocation, p. 358.

hence the rank of $\begin{pmatrix} A_{\mathcal{J}\mathcal{I}^*} \\ 0 \ldots 0 \end{pmatrix}$ is at least I. But the rank of $\begin{pmatrix} A_{\mathcal{J}\mathcal{I}^*} \\ 0 \ldots 0 \end{pmatrix}$ = the rank of $A_{\mathcal{J}\mathcal{I}^*}$. Therefore, the rank of $A_{\mathcal{J}\mathcal{I}^*}$ is I and $M_{\mathcal{J}\mathcal{I}^*}$ is non-singular.

2. The deletion of a constraint. Suppose ℓ intersects (but is not contained in) $\lambda_i = 0$ for $i > m_1$. We may assume $i = i_I$ and that

$$A_{\mathcal{J}\mathcal{I}} = \begin{pmatrix} A_{\mathcal{J}^*\mathcal{I}} \\ \alpha' \end{pmatrix}. \tag{45}$$

$A_{\mathcal{J}\mathcal{I}}$ has rank I, $A_{\mathcal{J}^*\mathcal{I}}$ has rank I-1; therefore $M_{\mathcal{J}^*\mathcal{I}}$ is non-singular.

3. Addition of a variable. Continuing the conventions used above, if $A_{\mathcal{J}\mathcal{I}}$ has rank I, so has $A_{\mathcal{J}\mathcal{I}^*} = (A_{\mathcal{J}\mathcal{I}}\ \alpha)$. Therefore $M_{\mathcal{J}\mathcal{I}^*}$ is non-singular.

4. Addition of a constraint. If ℓ intersects but is not contained in the plane $\xi_i = 0$, $i > m_1$ then

$$\begin{pmatrix} C_{\mathcal{J}\mathcal{I}} & A_{\mathcal{J}\mathcal{I}} & \mu_{j_1} \\ & & \vdots \\ & & \mu_{j_J} \\ A_{\mathcal{J}\mathcal{I}} & O & O \\ \alpha' & O & O \end{pmatrix} = \widetilde{M} \tag{46}$$

(where α' is the row of coefficients of the new constraint) is non-singular. Therefore $\begin{pmatrix} A_{\mathcal{J}\mathcal{I}} \\ \alpha \end{pmatrix} = A_{\mathcal{J}^*\mathcal{I}}$ has rank I + 1 and $M_{\mathcal{J}^*\mathcal{I}}$ is non-singular.

The tracing out of the efficient set is simplified if certain "accidents" do not occur. These accidents are described in the following "non-degeneracy" conditions. The next section of this paper presents a computing procedure for deriving the set of efficient points when all non-degeneracy conditions hold. In Sections 7-10 these conditions are relaxed.

CONDITION 1. On no critical line do we have

$$v_k = \alpha_{vk} + \beta_{vk} \lambda_E = 0 \quad \text{for } k \notin \mathcal{K}.$$

CONDITION 2. On any given critical line ℓ we do not have

$$\frac{-\alpha_{vk_1}}{\beta_{vk_1}} = \frac{-\alpha_{vk_2}}{\beta_{vk_2}} \quad \text{for any } k_1 \neq k_2 \text{ with } \beta_{vk_1} \neq 0,\ \beta_{vk_2} \neq 0.$$

CONDITION 3. E is bounded from above in $\widetilde{S}$.

We will let L_E stand for "the linear programming problem of maximizing E subject to constraints (1), (2), and (3)."

CONDITION 4. L_E has a unique non-degenerate solution.

Condition 4 implies condition 3.

6. THE ALGORITHM UNDER CONDITIONS 1 THROUGH 4

We now assume that conditions 1 through 4 are satisfied. Condition 4 implies that the optimum solution to L_E has:[14]

(a) Exactly m variables X_j and ξ_i are not at their lower extreme;

(b) $A_{\mathcal{I}(1)\,\mathcal{J}(1)}$ (where $\mathcal{I}^{(1)}$ includes all i with $\xi_i = 0$ and $\mathcal{J}^{(1)}$ includes all j with X_j not at its lower limit) has rank equal to the number of its rows;

(c) There exists "prices" p_i and "profitabilities" δ_j such that

$$p_i > 0 \quad \text{if } \xi_i = 0 \quad \text{for } i = m_1+1, \ldots, m, \tag{47}$$

$$p_i = 0 \quad \text{if } \xi_i > 0 \quad \text{for } i = m_1+1, \ldots, m, \tag{48}$$

$$\delta_j = \sum_i a_{ij} p_i + \mu_j, \tag{49}$$

$$\delta_j = 0 \quad \text{for } j = N_1+1, \ldots, N \text{ and for } X_j > 0 \quad j \le N_1, \tag{50}$$

$$\delta_j < 0 \quad \text{for } X_j = 0 \quad j \le N_1. \tag{51}$$

The matrix

$$M_{(1)} = \begin{pmatrix} C_{\mathcal{J}(1)\,\mathcal{J}(1)} & A'_{\mathcal{I}(1)\,\mathcal{J}(1)} \\ A_{\mathcal{I}(1)\,\mathcal{J}(1)} & O \end{pmatrix} \tag{52}$$

is non-singular and thus defines a critical line $\ell^{(1)}$ along which

$$M_{(1)} \begin{pmatrix} X_{\mathcal{J}(1)} \\ -\lambda_{\mathcal{I}(1)} \end{pmatrix} = \begin{pmatrix} O \\ B_{\mathcal{I}(1)} \end{pmatrix} + \begin{pmatrix} \mu_{\mathcal{J}(1)} \\ O \end{pmatrix} \lambda_E. \tag{53}$$

Since $-A'_{\mathcal{I}(1)\,\mathcal{J}(1)} \cdot p_{\mathcal{I}(1)} = \mu_{\mathcal{J}(1)}$, if $X^0_{\mathcal{J}(1)}\ \lambda^0_{\mathcal{I}(1)}$ satisfy (53) for $\lambda_E = 0$, then

$$M_1 \begin{pmatrix} X^0_{\mathcal{J}(1)} \\ -\lambda^0_{\mathcal{I}(1)} - p_{\mathcal{I}(1)}\,\lambda_E \end{pmatrix} = \begin{pmatrix} O \\ B_{\mathcal{I}(1)} \end{pmatrix} + \begin{pmatrix} \mu_{\mathcal{J}(1)} \\ O \end{pmatrix} \lambda_E \tag{54}$$

for all λ_E. Thus $\ell^{(1)}$ has

---

[14] The following are corollaries of the basis and pricing theorems of linear programming. See, e.g., George B. Dantzig, Alex Orden, Philip Wolfe, "The Generalized Simplex Method for Minimizing a Linear Form under Linear Inequality Restraints," Pacific Journal of Mathematics, Vol. 5, No. 2, June 1955.

$$
\begin{aligned}
\mathbf{X}_{\mathcal{J}(1)} &= \mathbf{X}^0_{\mathcal{J}(1)} \\
\lambda_{\mathcal{A}(1)} &= \lambda^0_{\mathcal{A}(1)} + p_{\mathcal{A}(1)}\,\lambda_E
\end{aligned}
\tag{55}
$$

for all λ_E. From (47) it follows that for sufficiently large λ_E

$$
\begin{aligned}
&\lambda_i > 0 \quad \text{for } i > m_1 \text{ and in } \mathcal{A}\,. \\
&\eta_j = \Sigma\sigma_{jh}\,\mathbf{X}^0_h - (\Sigma a_{ij}\,\lambda_i + \mu_j\lambda_E) \\
&\quad = \Sigma\sigma_{jh}\,\mathbf{X}^0_h - \Sigma a_{ij}\,\lambda^0_i - (\Sigma a_{ij}\,p_i + \mu_j)\lambda_E\,.
\end{aligned}
\tag{56}
$$

Equation (51) implies that for sufficiently large λ_E, $\eta_j > 0$ for $j \notin \mathcal{J}$. Thus for sufficiently large λ_E, $\mathcal{E}^{(1)}$ satisfies inequalities (40).

Let $\lambda_E^{(1)}$ be the largest value of λ_E at which $\mathcal{E}^{(1)}$ intersects a plane $\eta_j = 0$ for $j \notin \mathcal{J}$ or $\lambda_i = 0$ for $i \in \mathcal{A}$. (The $\mathbf{X}$ and $\mathcal{E}$ do not vary along $\mathcal{E}^{(1)}$.) If $\lambda_E^{(1)} \leq 0$ then $\mathbf{X}^0$ gives minimum V as well as maximum E. Suppose $\lambda_E^{(1)} > 0$. Non-degeneracy condition 2 implies $\mathcal{E}^{(1)}$ intersects only one plane $\eta_j = 0$ or $\lambda_i = 0$ at $\lambda_E^{(1)}$. In the former case we add j to $\mathcal{J}$; in the latter case we delete i from $\mathcal{A}$, to form $\mathcal{A}^{(2)}$, $\mathcal{J}^{(2)}$. The new matrix $\mathbf{M}_{(2)} = \mathbf{M}_{\mathcal{A}(2),\mathcal{J}(2)}$ is non-singular and defines a critical line $\mathcal{E}^{(2)}$. Suppose for a moment that it was $\eta_{j_0} = 0$ which $\mathcal{E}^{(1)}$ intersected at $\lambda_E^{(1)}$. On $\mathcal{E}^{(2)}$ we have at $\lambda_E = \lambda_E^{(1)}$:

$$
\begin{aligned}
&\lambda_i > 0 \quad \text{for } i > m_1 \text{ and } \in \mathcal{A}^{(2)}\,, \\
&\eta_i > 0 \quad \text{for } j \notin \mathcal{J}^{(2)}\,, \\
&\varepsilon_i > 0 \quad \text{for } i \notin \mathcal{A}\,, \\
&\text{and } \begin{cases} \mathbf{X}_{j_0} = 0 \\ \mathbf{X}_j > 0 \quad \text{for all other } j \leq N_1 \text{ and } \in \mathcal{J}^{(2)}\,. \end{cases}
\end{aligned}
\tag{57}
$$

As always $\mathbf{X}_{j_0} = a + b\lambda_E$ along $\mathcal{E}^{(2)}$. Non-degeneracy condition 1 assures $b \neq 0$. If $b < 0$ the projection of $\mathcal{E}^{(2)}$ would be efficient for $\lambda^* \geq \lambda_E \geq \lambda_E^{(1)}$ where $\lambda^* > \lambda_E^{(1)}$. This is impossible.[15] Therefore $b > 0$ and $\mathcal{E}^{(2)}$ is efficient for $\lambda_E^{(1)} \geq \lambda_E \geq \lambda_E^{(2)}$ where $\lambda_E^{(1)} > \lambda_E^{(2)}$. Similar remarks would apply if $\mathcal{E}^{(1)}$ first intersected $\lambda_{i_0} = 0$ and i_0 was deleted from $\mathcal{A}$.

---

[15] Since $b \neq 0$ the X-projection of the critical line is a line rather than a point. Along this line E increases with λ_E. If $\lambda_E > \lambda_E^{(1)}$ were feasible, then $E > E^{(1)} = \max E$ would be obtainable, which is impossible.

$\lambda_E^{(2)}$ is the highest value of $\lambda_E < \lambda_E^{(1)}$ at which $\ell^{(2)}$ intersects a plane $v_k = 0$ for $k = 1, \ldots, K$. If this is an η_j we again add a j to $\mathcal{J}$. If it is a λ_i we delete i from $\mathcal{I}$; if a ξ_i, we add i to $\mathcal{I}$; if an X_j, we delete j from $\mathcal{J}$. We form $M_{(3)}$ and $\ell_{(3)}$ accordingly and find $\lambda^{(3)} < \lambda^{(2)}$. This process is repeated until $\lambda_E = 0$ is reached. At each step (s) $M_{(s)}$ is non-singular and if v_{k_s} is the new variable (η, X, λ, or ξ) which is no longer constrained to be zero we have at $\lambda^{(s-1)}$.

(58)
$$v_k > 0 \quad \text{for } k \neq k_s \text{ and } \notin \mathcal{K},$$
$$v_{k_s} = 0.$$

By condition 1, $b_{vk_s} \neq 0$ along $\ell^{(s)}$. We argue below[16] that we cannot have $b_{vk_s} < 0$. So $b_{vk_s} > 0$ and $\ell^{(s)}$ is efficient for $\lambda_E^{(s-1)} \geq \lambda_E \geq \lambda_E^{(s)}$ where $\lambda_E^{(s-1)} > \lambda_E^{(s)}$. Since there are only a finite number of critical lines, and each can satisfy inequalities (40) for only one segment, $\lambda_E = 0$ is reached in a finite number of steps.

7. THE ALGORITHM UNDER CONDITIONS 3 AND 4

Let us now drop non-degeneracy conditions 1 and 2 but still assume conditions 3 and 4. We will use techniques analogous to the degeneracy-avoiding techniques of linear programming.[17]

For every number ε we define a new problem $P(\varepsilon)$ as follows:

$$\text{minimize } V(\varepsilon) = \Sigma\Sigma\sigma_{ij} X_i X_j + \Sigma \varepsilon^j X_j$$

subject to

(59)
$$\Sigma a_{ij} X_j = b_i + \varepsilon^{N+i}, \quad i = 1, \ldots, m_1,$$

(60)
$$\Sigma a_{ij} X_j \geq b_i + \varepsilon^{N+i}, \quad i = m_1+1, \ldots, m,$$

(61)
$$X_j \geq 0, \quad j = 1, \ldots, N_1.$$

---

[16] If v_{k_s} is an X_j or ξ_i, $b_{vk_s} < 0$ implies that there are two distinct points which minimize V for some $E > E^{(s-1)}$, which is impossible. This argument also applies if v_k is a λ_i or η_j unless the X-projection of the new critical line is a point. In the latter case we note (from the Kuhn and Tucker conditions) that an efficient point gives minimum $Q(\lambda_E) = V - \lambda_E E$ subject to (1), (2), (3). For fixed λ_E, $Q(\lambda_E)$ has a unique minimum. If $v_k < 0$ then two distinct points give minimum $Q(\lambda_E)$ for some $\lambda_E > \lambda_E^{(s-1)}$.

[17] In linear programming these techniques are generally not needed in practice. In quadratric programming arbitrary selection of $v_k = 0$ with $b_{vk} < 0$ to go into $\mathcal{K}$ may (or may not) prove adequate. In any case, the degeneracy-handling techniques are available if needed. See George Dantzig, "Application of the Simplex Method to a Transportation Problem," Activity Analysis of Production and Allocation, Tjalling C. Koopmans, ed.; A. Charnes, "Optimality and Degeneracy in Linear Programming," Econometrica, Vol. 20, No. 2, April, 1952, p. 160; and Dantzig, Orden, and Wolfe, Op. Cit.

391368 O - 56 - 9

For sufficiently small ε the unique, optimal basis of L_E is feasible and, since it still satisfies the pricing conditions, is optimal.

As we will see shortly for sufficiently small ε, $P(\varepsilon)$ satisfies non-degeneracy conditions (1) and (2). We will also see that for a sufficiently small ε^*, the sequence of indices $(\mathcal{J}\ \mathcal{I})^S$ associated with the critical lines $\ell^{(s)}$, until $\lambda_E = 0$ is reached, is the same for all $P(\varepsilon)$ for $\varepsilon^* \geq \varepsilon > 0$. If we change indices $(\mathcal{J}, \mathcal{I})$ in the same sequence as $P(\varepsilon)$ for small ε, if we let λ_E decrease along any critical line when it can without violating $v_k \geq 0$, until we reach $\lambda_E = 0$, then:

(a) We will pass through a finite number of index sets each associated with a non-singular $M_{\mathcal{J}\mathcal{I}}$, before we reach $\lambda_E = 0$.

(b) Since $v_k \geq 0$ is maintained we have the desired solution to the original problem.

Along any critical line of $P(\varepsilon)$ we have

$$\begin{pmatrix} X_{\mathcal{I}} \\ -\lambda_{\mathcal{J}} \end{pmatrix} = M_{\mathcal{J}\mathcal{I}}^{-1} \begin{pmatrix} O \\ B_{\mathcal{J}} \end{pmatrix} + M_{\mathcal{J}\mathcal{I}}^{-1} \begin{pmatrix} \mu_{\mathcal{I}} \\ O \end{pmatrix} \lambda_E + M_{\mathcal{J}\mathcal{I}}^{-1} \begin{pmatrix} \varepsilon^{j_1} \\ \vdots \\ \varepsilon^{N+i_I} \end{pmatrix} \tag{62}$$

or

$$X_{j_s} = \alpha_{X j_s} + \beta_{X j_s} \lambda_E + \sum_{h=1}^{I+J} m^{sh}\, \varepsilon^{f(h)} \tag{63}$$

where $f(1) = j_1, f(2) = j_2, \ldots, f(I+J) = N + i_I$

or

$$X_{j_s} = \alpha_{X j_s} + \beta_{X j_s} \lambda_E + p_{X j_s}(\varepsilon)\,. \tag{64}$$

Similarly

$$\lambda_{i_s} = \alpha_{\lambda i_s} + \beta_{\lambda i_s} \lambda_E - \sum_{h=1}^{I+J} m^{J+s,h}\, \varepsilon^{f(h)}$$

$$= \alpha_{\lambda i_s} + \beta_{\lambda i_s} \lambda_E + p_{\lambda i_s}(\varepsilon)\,.$$

$$\xi_i = \alpha_{\xi_i} + \beta_{\xi_i} \lambda_E + \sum_{s=1}^{J} a_{ij_s}\, p_{X j_s}(\varepsilon) + \varepsilon^{N+i} = \alpha_{\xi_i} + \beta_{\xi_i} \lambda_E + p_{\xi_i}(\varepsilon)\,. \tag{65}$$

$$\eta_j = \alpha_{\eta_j} + \beta_{\eta_j} \lambda_E + \sum_{s=1}^{J} \sigma_{jj_s}\, p_{X j_s}(\varepsilon) - \sum_{s=1}^{I} a_{i_s j}\, p_{\lambda i_s}(\varepsilon) + \varepsilon^{j} \tag{66}$$

$$= \alpha_{\eta_j} + \beta_{\eta_j} \lambda_E + p_{\eta_j}(\varepsilon)\,.$$

Consider the polynomials:

(67) $$p_{x_j}(\varepsilon) \quad \text{for} \quad j \in \mathcal{J},$$

(68) $$p_{\lambda i}(\varepsilon) \quad \text{for} \quad i \in \mathcal{I},$$

(69) $$p_{\xi i}(\varepsilon) \quad \text{for} \quad i \notin \mathcal{I},$$

(70) $$p_{\eta j}(\varepsilon) \quad \text{for} \quad j \notin \mathcal{J}.$$

None of the polynomials listed above have all zero coefficients, and no two have proportional coefficients. For each polynomial of (69) and (70) has a term with a coefficient of 1 which every other polynomial has with a coefficient of zero. This leaves only the possibilities that some polynomial of (67) or (68) has all zero coefficients or two of these polynomials have proportional coefficients. Both these possibilities imply that M^{-1} is singular and therefore are impossible.

Since $p_{v_k}(\varepsilon)$ has only a finite number of roots, for ε sufficiently small

$$p_{v_k}(\varepsilon) \neq 0 \quad \text{for} \quad k \notin \mathcal{K}.$$

Thus

(71) $$v_k = \alpha_{vk} + \beta_{vk}\,\lambda_E + p_{vk}(\varepsilon) \qquad -\infty < \lambda_E < \infty$$

cannot be identically zero for $k \notin \mathcal{K}$. The critical line intersects the plane $v_{k_1} = 0$ at

(72) $$\lambda'_E = \frac{-\alpha_{vk_1}}{\beta_{vk_1}} - \frac{p_{vk_1}(\varepsilon)}{\beta_{vk_1}}$$

and the plane

$$v_{k_2} = 0 \text{ at}$$

(73) $$\lambda''_E = \frac{-\alpha_{vk_2}}{\beta_{vk_2}} - \frac{p_{vk_2}(\varepsilon)}{\beta_{vk_2}}.$$

If, say,

(74) $$\frac{-\alpha_{vk_1}}{\beta_{vk_1}} > \frac{-\alpha_{vk_2}}{\beta_{vk_1}}$$

then for sufficiently small ε

$$\lambda'_E > \lambda''_E .$$

On the other hand, since

$$p_{vk_1}(\varepsilon) - p_{vk_2}(\varepsilon) = 0 \qquad k_1, k_2 \neq \text{※}$$

has a finite number of solutions, for sufficiently small ε

$$\lambda'_E \neq \lambda''_E$$

even if

$$\frac{-\alpha_{vk_1}}{\beta_{vk_1}} = \frac{-\alpha_{vk_2}}{\beta_{vk_2}} . \tag{75}$$

As $\varepsilon \longrightarrow 0$ the smallest power of ε dominates; i.e., if, say, $\dfrac{-p_{vk_1}(\varepsilon)}{\beta_{vk_1}}$ has an algebraically larger coefficient for the first power of ε, then $\lambda' > \lambda''$ as $\varepsilon \longrightarrow 0$. If both have the same coefficient of ε, then it is the coefficients of ε^2 that decide. And so on.

Since there are a finite number of critical lines and a finite number of planes $v_k = 0$, there is a single ε^* such that for $\varepsilon^* \geq \varepsilon > 0$.

$P(\varepsilon)$ satisfies non-degeneracy conditions (1) and (2); and the order of the index sets $(\mathcal{A}\,\mathcal{G})^s$ is the same for all such ε.

The m^{st} are needed for other purposes and are thus available for resolving degeneracy problems. The other coefficients of $p_{v_k}(\varepsilon)$ can be computed when needed.

8. THE ALGORITHM WHEN L_E IS DEGENERATE BUT UNIQUE

Suppose that the solution to L_E is degenerate in that one or more of the basis variables X_j or ξ_i is "accidentally" zero, but is unique in that $\delta_j < 0$ for all X_j not in the basis and $p_i > 0$ for all ξ_i not in the basis.

The constraints of L_E may be written as a system of equations including the ξ_i as variables:

$$B \begin{pmatrix} X \\ \xi \end{pmatrix} = b . \tag{76}$$

If $\widetilde{B}$ is the submatrix of optimal basis vectors and if $X_{\mathcal{G}}$ and $\xi_{\bar{\mathcal{A}}}$ are the optimal basis variables, then the optimal solution is given by

$$\begin{pmatrix} X^0_{\mathcal{G}} \\ \xi^0_{\bar{\mathcal{A}}} \end{pmatrix} = \widetilde{B}^{-1} b \tag{77}$$

while all other variables are zero. After we solve L_E we may modify it, forming $L_E(\varepsilon)$ as follows:

$$B \begin{pmatrix} X \\ \xi \end{pmatrix} = b + \tilde{B} \begin{pmatrix} \varepsilon \\ \varepsilon \\ \vdots \\ \varepsilon \end{pmatrix} \quad \text{for } \varepsilon > 0 \tag{78}$$

$$= b + \varepsilon \begin{pmatrix} r_1 \\ \vdots \\ r_m \end{pmatrix}$$

where r_i is the sum of the i^{th} row of $\tilde{B}$.

Then

$$\begin{pmatrix} X_{\mathcal{J}}(\varepsilon) \\ \xi_{\bar{\mathcal{J}}}(\varepsilon) \end{pmatrix} = \tilde{B}^{-1} b + \begin{pmatrix} \varepsilon \\ \vdots \\ \varepsilon \end{pmatrix}. \tag{79}$$

Thus the original optimal basis is still feasible and therefore uniquely optimal (since it still satisfies the pricing relationships). Also for $\varepsilon > 0$

$$X_j(\varepsilon) > 0 \qquad \text{for } j \in \mathcal{J},$$

$$\xi_i(\varepsilon) > 0 \qquad \text{for } i \in \bar{\mathcal{J}},$$

and

$$\begin{pmatrix} X_{\mathcal{J}}(\varepsilon) \\ \xi_{\bar{\mathcal{J}}}(\varepsilon) \end{pmatrix} \longrightarrow \begin{pmatrix} X^0_{\mathcal{J}} \\ \xi^0_{\bar{\mathcal{J}}} \end{pmatrix}$$

as $\varepsilon \longrightarrow 0$.

The procedures of the last section which apply when L_E has a unique and non-degenerate solution apply with essentially no modification if L_E has a unique but possibly degenerate solution, if we let $P(\varepsilon)$ be

$$\min V = \Sigma\Sigma \sigma_{ij} X_i X_j + \Sigma \varepsilon^{j+1} X_j$$

subject to

$$\Sigma a_{ij} X_j = b_i + r_i \varepsilon + \varepsilon^{N+i+1} \quad \text{for } i = 1, \ldots, m_1,$$

$$\Sigma a_{ij} X_j \geq b_i + r_i \varepsilon + \varepsilon^{N+i+1} \quad \text{for } i = m_1+1, \ldots, m. \tag{80}$$

The solution to $L_E(\varepsilon)$ is non-degenerate for sufficiently small ε. Along any critical line we now have

$$X_{j_s} = \alpha_{X j_s} + \beta_{X j_s} \lambda_E + \varepsilon\, p_{X j_s}(\varepsilon) + \left(\sum_{h=1}^{I} m^{s,\, h+J} r_{i_h} \right) \varepsilon$$

(81)

$$= \alpha_{X j_s} + \beta_{X j_s} \lambda_E + q_{X j_s}(\varepsilon) ;$$

$$\lambda_{i_s} = \alpha_{\lambda i_s} + \beta_{\lambda i_s} \lambda_E + \varepsilon\, p_{\lambda i_s}(\varepsilon) - \left(\sum_{h=1}^{I} m^{s+J,\, h+J} r_{i_h} \right) \varepsilon$$

(82)

$$= \alpha_{\lambda i_s} + \beta_{\lambda i_s} \lambda_E + q_{\lambda i_s}(\varepsilon) ;$$

$$\xi_i = \alpha_{\xi i} + \beta_{\xi i} \lambda_E + \sum_{s=1}^{J} a_{ij_s} q_{X j_s}(\varepsilon) + \varepsilon^{N+i+1}$$

(83)

$$= \alpha_{\xi i} + \beta_{\xi i} \lambda_E + q_{\xi i}(\varepsilon) ;$$

$$\eta_j = \alpha_{\eta j} + \beta_{\eta j} \lambda_E + \sum_{s=1}^{J} \sigma_{jj_s} q_{X j_s}(\varepsilon) - \sum_{s=1}^{I} a_{i_s j} q_{\lambda i_s}(\varepsilon) + \varepsilon^{j+1}$$

(84)

$$= \alpha_{\eta j} + \beta_{\eta j} \lambda_E + q_{\eta j}(\varepsilon)$$

where the $p_v(\varepsilon)$ are as defined in (63) through (66). Since no $p_v(\varepsilon)$ can have zero coefficients and no two can have proportional coefficients, the same is true of the $q_v(\varepsilon)$.

9. THE ALGORITHM WHEN L_E IS NOT UNIQUE

A non-degenerate optimal solution to L_E is unique if, and only if,

$$\delta_j < 0 \text{ for } X_j \text{ not in the basis}$$

(85)

$$p_i > 0 \text{ for } \xi_i \text{ not in the basis.}$$

If L_E has a degenerate solution and (85) does not hold, then either the solution is not unique or else only the optimal basis is not unique. If L_E does not have a unique solution we must find the point $\bar{X}$ which gives minimum V for $E = \bar{E} = \max E$. If only the optimal basis of L_E is not unique we still must decide on the $\Delta\xi$ of our first critical line. Both these problems will be resolved in the same manner. Our procedure may be considered as a special case of either approach 1 or approach 4 for minimizing a quadratic subject to linear constraints described in Section 12.

Let us create a new linear programming problem $L_F(\varepsilon)$ by adding a constraint to and modifying the form to be maximized in $L_E(\varepsilon)$. The equation we add is

$$\sum_{j=1}^{N} \mu_j X_j - \xi_E = \bar{E} + \left(\sum_{j \in \mathcal{J}} \mu_j - 1 \right) \varepsilon . \tag{86}$$

If we add ξ_E to the optimum basis variables of $L_E(\varepsilon)$ we have a feasible basis corresponding to a solution with

$$X_j = X_j^0 + \varepsilon \qquad j \in \mathcal{J},$$

$$\xi_i = \xi_i^0 + \varepsilon \qquad i \in \bar{\mathcal{A}}, \tag{87}$$

$$\xi_E = \varepsilon .$$

Next let us replace the objective function $E = \Sigma \mu X$ with a new one

$$F = \Sigma \nu_j X_j \tag{88}$$

such that the solution in (87) is the unique optimum of $L_F(\varepsilon)$. This may be done easily by assigning any values $p_i > 0$ to $i \notin \bar{\mathcal{A}}$, $p_i = 0$ for $i \in \bar{\mathcal{A}}$ as well as $p_E = 0$. Then choose any set for ν_j so that $\delta_j = 0$ for $j \in \mathcal{J}$ and $\delta_j < 0$ for $j \notin \mathcal{J}$. Since $L_F(\varepsilon)$ has a unique non-degenerate solution we may use methods already described to trace out the set of points which give minimum $V(\varepsilon)$ for given F until $\lambda_F = 0$. If only a few bases are feasible for $L_F(\varepsilon)$, i.e., if not too many bases are optimal for $L_E(\varepsilon)$, $\lambda_F = 0$ will be reached quickly. At $\lambda_F = 0$ either $\lambda_E = 0$ or $\lambda_E > 0$. In the former case we have arrived at a point with minimum V and maximum E. In the latter case we have $\bar{X}$ and are ready to trace out the set of efficient X's. From this point on we let $\lambda_F = 0$, i.e., we ignore F completely. Since at $\lambda_F = 0$, $v_k > 0$ for all $k \notin \mathcal{K}$ we may reduce λ_E until we intersect a plane $v_k = 0$ and continue as in Section 8.

10. THE ALGORITHM, WHEN CONDITION 3 DOES NOT HOLD

If E is unbounded procedure 4 of Section 12 can be used to find the point $\underline{X}$ with minimum V. The efficient set can then be traced out in the direction of increasing λ_E. Since there are only a finite number of critical lines and each critical line is efficient at most once the efficient set is traced out in a finite number of steps.

11. THE SET OF EFFICIENT E, V COMBINATIONS

Once the set of efficient X's is found the set of efficient E, V combinations can be obtained easily. The critical line of a subspace in which more than one value of E is obtainable may be expressed as the solution to

$$\Sigma \sigma_{jk} X_k + \Sigma(-\lambda_i) a_{ij} + (-\lambda_E)\mu_j = 0, \qquad j \in \mathcal{J}, \tag{89}$$

$$\Sigma a_{ij} X_j = b_i, \qquad i \in \mathcal{A}, \tag{90}$$

$$\Sigma \mu_j X_j = E \tag{91}$$

$$\text{for } -\infty < E < +\infty .$$

If we let N^{-1} be the inverse of the matrix N in (89), (90), (91) we have

$$\begin{pmatrix} X \\ -\lambda \end{pmatrix} = N^{-1} \begin{pmatrix} 0 \\ b \\ E \end{pmatrix} \tag{92}$$

$$V = (X', -\lambda') \begin{pmatrix} C & O \\ O & O \end{pmatrix} \begin{pmatrix} X \\ -\lambda \end{pmatrix} = (0', b', E) N^{-1} C N^{-1} \begin{pmatrix} 0 \\ b \\ E \end{pmatrix} \tag{93}$$

from which it follows that along any such critical line V and E are related by a formula of the form

$$V = a + bE + CE^2 . \tag{94}$$

Thus the set of efficient E, V combinations is piecewise parabolic. We know, or can easily get, the values of E and dV/dE at the end points of each of the pieces. We can also evaluate V at $\bar{X}$.[18] Knowing V at one value of E and $dV/dE = b + 2CE$ at two values of E we can solve for the a, b, and c in (94) for the segment from $\bar{E}$ to $\bar{E} - \varepsilon_1$. Having a, b, and c we can evaluate V at $\bar{E} - \varepsilon_1$ by means of (94). This provides us with the value of V at one value of E on the segment which is efficient from $E - \varepsilon_1$ to $E - \varepsilon_2$. This, combined with the values of dV/dE at two values of E, allows us to obtain the a, b, c of (94) for this next segment—and so on until we trace out the set of E, V combinations.

12. MINIMIZING A QUADRATIC

One of the "by-products" of the calculation of efficient sets is the point at which V is a minimum, i.e., where $\lambda_E = 0$. The computing procedures described in Sections 6 through 10 are analogous to the simplex method of linear programming (as contrasted with the "gradient methods" that have been suggested for both linear and non-linear programming). Both the procedure described in the preceding section—considered as a way of getting to min V—and the simplex method require a finite number of iterations, each iteration typically taking a "jump" to a new point which is superior to the old. Each iteration makes use of the inverse of a matrix which is a "slight" modification of the matrix of the previous iteration. The success of the simplex method in linear programming suggests that it may be desirable to use a variant of the "critical line" method in the quadratic case.

Our problem then is to minimize a quadratic

$$V = \Sigma\Sigma\, \sigma_{ij} X_i X_j$$

subject to constraints (1), (2) and (3). We wish to translate this into a problem of tracing out an efficient set. This may be done in several ways.

1. An arbitrary set of μ_j can be selected and the efficient set traced out until $\lambda_E = 0$. The μ_j should be selected so that the "artificial" E has a unique maximum.

2. An equality, say,

$$\Sigma a_{1j} X_j = b_1$$

[18] If $\bar{X}$ does not exist we can evaluate V at $\underline{X}$ and use the same process "in reverse."

can be eliminated from (1). E can be defined as

$$E = \Sigma a_{1j} X_j$$

and the critical set traced out until $E = b_1$. If $E = b_1$ is reached before $\lambda_E = 0$ the computing procedures of the last section must be continued into the region of $\lambda_E < 0$. While the points thus generated will not be efficient—for they do not give max E for given V—they do give min V for given E. In particular, they will arrive at the point of min V for

$$E = \Sigma a_{1j} X_j = b_1 .$$

3. An inequality, say,

$$\Sigma a_{mj} X_j \geq b_m$$

can be eliminated from (2). E can be defined as

$$E = \Sigma a_{mj} X_j .$$

The efficient set is traced out until either $E = b_m$ or else $\lambda_E = 0$. If the former happens first the constraint is effective; if the latter happens first the constraint is ineffective. In either case, the point associated with the first of these to occur gives min V subject to (1), (2), and (3).

4. An initial guess $X_1^0, \ldots, X_N^0$ which satisfies (1), (2), and (3) can be made and μ_j defined so that, given these μ_j, X^0 is efficient. The efficient set can then be traced out until $\lambda_E = 0$. To choose μ_j so that X^0 is efficient, choose arbitrary positive values of λ_i ($i \in \mathfrak{A}$) and λ_E. Then choose μ_j so that

$$\eta_j = 0 \quad \text{for } X_j \text{ not at its lower bound,}$$

$$\eta_j > 0 \quad \text{for } X_j \text{ at its lower bound.}$$

If X^0 is in the same subspace as the optimal solution, the latter is reached in one iteration.

---

[19] It has been found recently that the strict convexity assumption can be relaxed. The procedures described in this paper apply, without modification, to the homogeneous quadratic whenever (σ_{ij}) is positive semi-definite. The non-homogeneous quadratic requires a slight modification of procedure for the general semi-definite (σ_{ij}). (Footnote added in proof.)

* * *

NEWS AND MEMORANDA

Professor A. Charnes of the Mathematics Department of Carnegie Institute of Technology has been appointed Professor of Mathematics and Industrial Management at Purdue University. This is a joint appointment between the Mathematics Department and the new School of Industrial Engineering and Management. He has also been named Director of Research in the latter.

Dr. W. H. Marlow was appointed Principal Investigator of The George Washington University Logistics Research Project on February 1, 1956. During the previous year he was associated with Numerical Analysis Research, Department of Mathematics, University of California, Los Angeles.

Dr. M. E. Rose of the Office of Naval Research has been appointed Managing Editor of the Naval Research Logistics Quarterly.

The following appointments to Associate Editor are announced:

Dr. A. J. Hoffman, National Bureau of Standards
Dr. S. Karlin, California Institute of Technology
Dr. W. H. Marlow, George Washington University Logistics Research Project
Dr. M. Wood, Office of the Secretary of Defense
Dr. M. A. Geisler, The RAND Corporation
Captain P. L. Folsom, USN, Staff, Naval War College

Appreciation is expressed to the following individuals for their valuable services as Associate Editors of this journal during their terms of office: Prof. D. Blackwell, Capt. R. M. Bowstrom, Dr. G. B. Dantzig, Prof. M. M. Flood, and Cdr. H. Strock.

The Second Annual Statistical Engineering Symposium sponsored by the Chemical Corps Engineering Agency will be held at the Army Chemical Center, Maryland on April 26 and 27, 1956.

A Symposium for Management of Industrial Applications of Analog Computers, sponsored by the Midwest Research Institute will be held on April 10 and 11, 1956, at the Hotel Phillips, Kansas City, Missouri.

The Fourth Annual Meeting of the Operations Research Society of America will be held in Washington, D. C., on May 10 and 11, 1956, at the Sheraton-Park Hotel. The program will consist of invited addresses and sessions for both invited and contributed papers. President LeRoy A. Brothers will give his retiring address at a banquet on Thursday evening, May 10.

* * *

RECENT PUBLICATIONS

THE DIRECTION OF WAR by Air Vice-Marshall Edward J. Kingston-McCloughry, R.A.F. Frederick A. Praeger, Inc. New York, 261 pp.

This book is subtitled "A Critique of the Political Direction and High Command in War" and according to the author is the first such essay written on the subject from an inter-Service and inter-Allied approach. Air Vice-Marshall Kingston-McCloughry writes out of his experiences in a wide variety of high command positions. The student of organization theory will find this volume particularly useful with respect to the organization of high command and the influence of personality on any particular organization. The author feels that it is important to evaluate the systems and personalities involved in high command because of the appalling losses in manhood and material that are caused by mistaken political direction and High Command. By the very nature of things, such high level staffs do not write their own accounts for publication critically and impartially. This, then, is the author's attempt to reveal the factors lying behind high command decisions and the efficacy of various organizations and personalities to make proper decisions.

MERCHANT SHIPPING AND THE DEMANDS OF WAR by C. B. A. Behrens. London, Her Majesty's Stationery Office, 1955, 494 pp.

One of the Official British Civil Histories of the Second World War, this book is concerned with the problem of planning for war shipping requirements in the United Kingdom. It presents a thoroughly documented and detailed account of how successful the control of shipping in war-time was as the result of collaboration between the Ministry of War Transport, civilian claimants, and the Services. In short, the text ably portrays that though the organization that controls merchant ships can do nothing directly to win a war, it can very easily cause one to be lost. Despite the vast amount of detail which is magnified by numerous appendices, statistical in nature, this is a story that needs to be told. Its value to the student lies in the collecting of a great amount of research material into one volume for ready, though not too readable, assimilation. The author makes some interesting observations concerning British dealings with the Americans and, more particularly, some of the "shortcomings" of American planning in the War Shipping Administration. The text is supplemented with pictures and illustrations, and the topic and chapter headings make it useful as a reference work.

WORLD COMMERCE AND GOVERNMENTS: TRENDS AND OUTLOOK by W. S. & E. S. Woytinsky. New York, The Twentieth Century Fund, 1955, 907 pp.

A sequel to WORLD POPULATION AND PRODUCTION, this is a valuable reference work. It is divided into three main parts: Trade, Transportation, and Governments. The consideration of each subject begins with an historical development which is carried to the

present. Containing many tables of well-substantiated comparative figures, it is not only a valuable source of statistical information but contains the background information required for its proper understanding.

CAPITAL FORMATION AND ECONOMIC GROWTH edited by Moses Abramovitz. Published for the National Bureau of Economic Research by Princeton University Press, 1955, pp. xiii + 677.

This volume contains the papers presented at the Conference on Capital Formation and Economic Growth which met in New York, November 1953, under the auspices of the Universities-National Bureau Committee for Economic Research. The volume also includes discussion papers prepared at the invitation of the Conference Committee and contributions submitted by members of the Conference. The authors include economists, historians, and a sociologist. Simon Kuznets and Raymond Goldsmith examine the sources and channels of finance in capitalist nations. Gregory Grossman and F. D. Holzman analyze the experience of the Soviet Union. Bert Hoselitz and Thomas Cochran consider the effects of entrepreneurship and business organizations on capital formation in Britain, France, and the United States. Henry Aubrey and Marion Levy discuss the importance of the same factors in underdeveloped areas. Abbott Payson Usher and W. Rupert Maclaurin relate invention and innovation to investment, while Adolph Lowe uses a series of models to analyze the physical composition of investment required for balanced growth. W. W. Rostow sums up the contributions and suggests further points for study. In addition, Moses Abramovitz has written an introduction to the study of capital formation.

THE ELEMENTS OF PROBABILITY THEORY AND SOME OF ITS APPLICATIONS by Harald Cramer. New York: John Wiley and Sons, Inc., 1955, 281 pp.

This book is a revised and extended translation of a Swedish textbook first published in 1926 and rewritten in 1949. The book is by the author of the famous work entitled MATHEMATICAL METHODS OF STATISTICS, published by Princeton University Press in 1946. The work is of an elementary level and requires that the reader have only some working knowledge of analytical geometry, calculus, and algebra. The author believes that the book could be read without difficulty by a student at the junior level in an ordinary American college. There are three parts. The first deals with foundations, giving the principal theorems for probability and definitions of terms used in the later part. The second part discusses random variables and probability distribution, including the chi square distribution. The third part brings applications mostly to statistics, in particular to sampling distribution and the problems of statistical inference. Theorems are proved as far as possible, but where the proofs would be too difficult the reader is referred to the more advanced literature which is always carefully indicated. Many examples and exercises are provided. There are long sections which are non-mathematical and which could serve for those readers who do not wish to work their way through all parts of the book, yet they would obtain a reasonable understanding of the principal notions of probability theory.

HARMONIC ANALYSIS AND THE THEORY OF PROBABILITY by Salomon Bochner. University of California Press, 1955, 176 pp.

In this volume the author ties together and expands upon a wide variety of advanced topics in Fourier analysis with finitely and infinitely many variables and in the theory of probability, exhibiting some of the very intimate connections between the two fields. Many of the author's results of recent years have been expanded upon and their interrelations further explored; some results appear for the first time. The topics taken up include: Kernel-approximation theory in finitely many variables; Fourier expansions in finite or infinite dimensions; closure properties of the Fourier transform; Laplace and Mellin transforms and the subordination of certain stochastic processes; characteristic functionals; and an analysis of expansions and convergence in stochastic processes.

INTRODUCTION TO MATHEMATICAL STATISTICS by Paul G. Hoel. New York: John Wiley and Sons, Inc., 2nd edition, 1954, pp. xi + 331.

This textbook on mathematical statistics is a rather extensive revision of the first edition of the author's well-known book. The second edition places more emphasis on the basic theory of statistics, but the material is presented so that the reader needs only a sound background in elementary calculus to understand the theoretical developments. A new chapter on probability has been added and other chapters have been completely rewritten or enlarged. The chapters on correlation and regression have been rewritten to place greater emphasis on regression. The material on the analysis of variance has been rewritten and expanded considerably. The chapter on nonparametric methods also has been expanded. Finally, a large number of exercises have been added to the chapters.

* * *

U. S. GOVERNMENT PRINTING OFFICE : 1956 O - 391368